QUANTIFYING THE IMPORTANCE OF THE RARE BIOSPHERE FOR MICROBIAL COMMUNITY ADAPTATION TO ORGANIC POLLUTANTS IN A FRESHWATER ECOSYSTEM

A Thesis
Presented to
The Academic Faculty

by

Yuanqi Wang

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in the
School of Civil and Environmental Engineering

Georgia Institute of Technology
May 2016

QUANTIFYING THE IMPORTANCE OF THE RARE BIOSPHERE FOR MICROBIAL COMMUNITY ADAPTATION TO ORGANIC POLLUTANTS IN A FRESHWATER ECOSYSTEM

Approved by:

Dr. Kostas Konstantinidis, Advisor
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Jim C. Spain
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Spyros Pavlostathis
School of Civil and Environmental Engineering
Georgia Institute of Technology

Date Approved: 04/29/2016

ACKNOWLEDGEMENTS

I would like to thank my parents first. They provided the opportunity for me to study in the United States. I also would like to thank my advisor Dr. Kostas Konstantinidis. I could not have accomplished my research without his support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Microbial community analysis frequently focuses on abundant organisms, but natural communities commonly harbor thousands of low abundance, 'rare' organisms. The importance of this 'rare biosphere' for microbial community adaptation to environmental perturbations remains speculative. We tested whether rare species respond to changing environmental conditions by establishing 20 liter, planktonic mesocosms with water from Lake Lanier (Georgia, USA), and perturbing them with organic compounds that are rarely detected in the lake, such as 2,4-dichlorophenoxyacetic acid (2,4-D), 4-nitrophenol (4-NP), and caffeine. The populations of the degraders of these compounds were initially below detection limit of qPCR or metagenomic sequencing methods (i.e., <0.001% of total community), but increased substantially in abundance after perturbation. All replicated 2,4-D mesocosms exhibited distinct degradation profiles, presumably due to the response of the rare biosphere (e.g., not all species were present in all inocula) and/or stochastic processes in the activation of rare species and genes. To obtain further insights into the latter, we sequenced several 2,4D-degrading isolates recovered from the mecososms, and assessed their genomes against time-series metagenomic datasets from the 2,4-D replicated mesocosms. We found distinct, co-occurring alleles of degradation genes, encoded frequently on transmissible plasmids, and distinct species dominated the post-enrichment datasets from each mesocosm. Collectively, these results supported the hypothesis that the rare biosphere can serve as a

genetic reservoir that enables community adaptation to changing environmental conditions, and provided insights into the size of the pool of rare genes and species.

CHAPTER 1

INTRODUCTION

Extant biodiversity is recognized as telling the evolutionary history of life while also providing an evolutionary scaffold for the future. Consequently, one of the great challenges for environmental microbiology, and the natural sciences, is to better understand how the inventory of biodiversity determines the evolutionary path(s) that will shape the future. Microbial communities in terrestrial or aquatic habitats are typically composed of hundreds, if not thousands, of distinct species (Torsvik, Goksoyr et al. 1990, Whitman, Coleman et al. 1998, Curtis, Sloan et al. 2002), each of which typically makes up a rather small fraction of the total community, and encodes hundreds of species-specific genes of unknown function (Nelson, Paulsen et al. 2000, Konstantinidis and Tiedje 2005). The pool of low-abundance species has been termed the "rare biosphere" (Sogin, Morrison et al. 2006), although the definition of "rare" is typically based on arbitrary cut-offs in abundance, e.g., < 0.1% of the total community (Pedrós-Alió 2012). A quantitative understanding of the contributions of this rare biosphere to the process of community adaptation within periods of time that are relevant for human activities (e.g., days to months) remain elusive (Konstantinidis, Ramette et al. 2006).

It has been suggested that most bacterial species found in a given habitat represent important and ancient players of the indigenous community and contribute substantially to community function and resilience, for instance by serving as a source of genomic innovation through the species-specific metabolic diversity they harbor (Sogin, Morrison

et al. 2006, Campbell, Yu et al. 2011). On the other hand, it is also thought that the majority of this species diversity represents a sort of "biological detritus" accumulating from a combination of very efficient microbial dispersal and slow decay kinetics of individual cells (Falkowski, Fenchel et al. 2008). Furthermore, high biodiversity observed in several habitats results, at least in part, from sequencing artifacts (Huse, Welch et al. 2010) (Quince, Lanzén et al. 2009) or free DNA released from dead cells (Stoeck and Epstein 2009). These two contrasting views are not necessarily mutually exclusive; for instance, while the majority of rare species may never contribute to community function and adaptation, it can be hypothesized that a fraction of them does infrequently contribute to adaptation, depending on selection pressures from the changing environment. Obtaining a quantitative view of the number of species (or sequence-based operational taxonomic units) representing each of these two perspectives within representative ecosystems on Earth will lead to a better understanding of the extant biodiversity. Although few examples have been reported in which low-abundance species were shown to carry out major ecological roles [e.g., (Musat, Halm et al. 2008, Shade, Jones et al. 2014)], our understanding of the frequency by which low-abundance species and genes participate in and the mechanisms used for community adaptation to perturbations is far from complete. Obtaining a quantitative understanding of the process will require following the temporal changes in composition and activity of natural microbial communities after perturbations in well-replicated experiments/samples using high resolution approaches.

To provide novel insights into the issues discussed above and test prevailing hypotheses about the ecological role of the rare biosphere, we set up parallel mesocosms

in the laboratory derived from a well-characterized planktonic community from Lake Lanier (Georgia, USA), and challenged them with relatively common organic compounds that were not detectable in the lake at the time of sampling (using compounds that are abundant in the lake would not have been as informative about the rare biosphere since they typically select for abundant community members). The mesocosms were sampled repeatedly for metagenomic analysis to follow the evolution of the bacterial communities and identify which populations responded to the treatment (e.g., became enriched or depleted over time). Parallel-unamended mesocosms served as controls and references. The results allowed us to rigorously test the hypothesis that low abundance members, as opposed to genes of the major components of the community, provided the metabolic diversity that enabled the community to respond to these changing conditions. Further, the data provided insights into the variability and redundancy of the responding low abundance populations in different samples from this freshwater ecosystem, as well as the genetic diversity of the catabolic genes underlying the biodegradation of the added compounds.

CHAPTER 2

MATERIALS AND METHODS

Site description and sampling

Water samples were collected on December 15, 2013 from Lake Lanier, GA (34°N 15′43″, 83°W 57′7″). This seasonally stratified lake is a manmade reservoir located in the northern portion of the state of Georgia, and serves as the primary drinking water resource for the Atlanta metropolitan area. The primary inflows are the Chattahoochee River and Chestatee River, and the surface area is approximately 150 km^2.

A horizontal sampler (Wildco Instruments, Yulee, FL, USA) was used to collect samples of planktonic microbial communities at a depth of 5 m. This depth was chosen because it represents the well-oxygenated and highly productive layer of the water column. Moreover, this depth is within the epilimnion layer (the top-most surface layer), which is fairly uniform in temperature and has higher dissolved oxygen concentration.

Mesocosm experiment set-up, sampling strategy, and DNA extraction

2,4-Dichlorophenoxyacetic acid (2,4-D), a widely used herbicide, 1,3,7-trimethyluric acid (caffeine), and 4-nitrophenol (4-NP), a precursor compound of several chemicals such as fungicides and a decomposition product of pesticides, were used to perturb the community. The initial added concentrations of 2,4-D, caffeine, and 4-NP were 40 μM, 100 μM, and 150 μM, respectively. These three chemicals were chosen because 1) their concentrations in the lake were below detection limit (<5 μM) of high-

performance liquid chromatography (HPLC), 2) biodegradation pathways and the underlying genes are known, which facilitated analysis of the time-series metagenomes, and 3) these organic molecules should exert selection pressure on populations by supporting the growth of degrading bacteria.

The 4-NP and caffeine mesocosm enrichments were initiated on December 15, 2013 and the 2,4-D mesocosm enrichment started on December 26, 2013. Glass water bottles (20L) used for the mesocosms, were rinsed with 10% fresh hydrochloric acid, followed by three washes with distilled water then sterilized by autoclaving prior to addition of 18L of lake water. Mesocosms were established in triplicate and incubated at room temperature with gentle mixing in the dark. Two controls were established for each compound: one abiotic control containing sterilized lake water (autoclaved for 180 min, using a fluid cycle) supplemented with each of the three chemicals; and a bottle-effect (BE) control, non-sterilized lake water incubated under the same conditions without the addition of the compound. HPLC was used to monitor the concentrations of the compounds over the incubation period. A filtration system, essentially as described previously [(Oh, Caro-Quintero et al. 2011) and Supplementary Figure S1] was assembled to collect samples for DNA sequencing. Briefly, a total of 1L or 5 L of water was used in each sample (Figure 1), pre-filtered through AP filters (~5 µm; Millipore, Billerica, MA, USA) and GF/A filters (~1.6 µm; Whatman, Little Chalfont, UK), and cells were collected on Sterivex filters (~0.22 µm; Millipore) using a peristaltic pump. Mesocosm biomass was collected at specific time points based on the disappearance profile of the added compound; T = 0, 2, 4, and 5 in 2,4-D mesocosm I (sample IDs: E1-

T0, E1-T2, E1-T4, and E1-T5), at T = 3 in 2,4-D mesocosm III (E3-T3), and at T = 4 in the BE control (BE-T4) (Figure 1). Supplementary Table 1 shows the 2,4-D mesocosm samples taken, including the number of Sterivex filters used, sample volume for each Sterivex filter, and total sample volume collected at each time point. Five liter samples were taken from the 4-NP mesocosm I was sampled at T = 2 (4NP-E1) and caffeine mesocosm I , III and BE control at T = 2 (Caff-E1, Caff-E3 and Caff-BE) (Figure 2). Filters were stored at -80°C until used for DNA extraction. A phenol-chloroform DNA extraction protocol was used (Oh, Caro-Quintero et al. 2011) with minor modifications. First, Sterivex filter units were opened and filters were removed and split into two equal pieces aseptically, before being placed in a 2 mL screw-top, O-ringed tubes on ice. Lysis buffer (40 mM EDTA, 50 mM Tris-HCl, and 0.73 M sucrose) and 1.15 mg/mL lysozyme were added. The mixture was incubated at 37°C for 30 minutes. Samples were subsequently incubated with 1% SDS (sodium dodecyl sulfate), 0.65 mg/mL Proteinase K, and 200 µg/mL RNase at 55°C for 2 hours in a rotating hybridization oven. DNA was extracted from the lysate with phenol and chloroform, precipitated with 70% ethanol, and eluted in Tris-EDTA buffer. All mesocosm metagenome libraries for Illumina sequencing were prepared using the Illumina Nextera XT DNA library prep kit as described by the manufacturer. For 2,4-D mesocosms, DNA libraries were sequenced on an Illumina HiSEQ 2500 instrument (Georgia Institute of Technology) for 300 cycles (2 x 150 bp paired end run). DNA libraries from Caffeine and 4N-P mesocosm samples-were sequenced on an in-house Illumina MiSeq instrument for 500 cycles (2 x 250 bp paired end run).

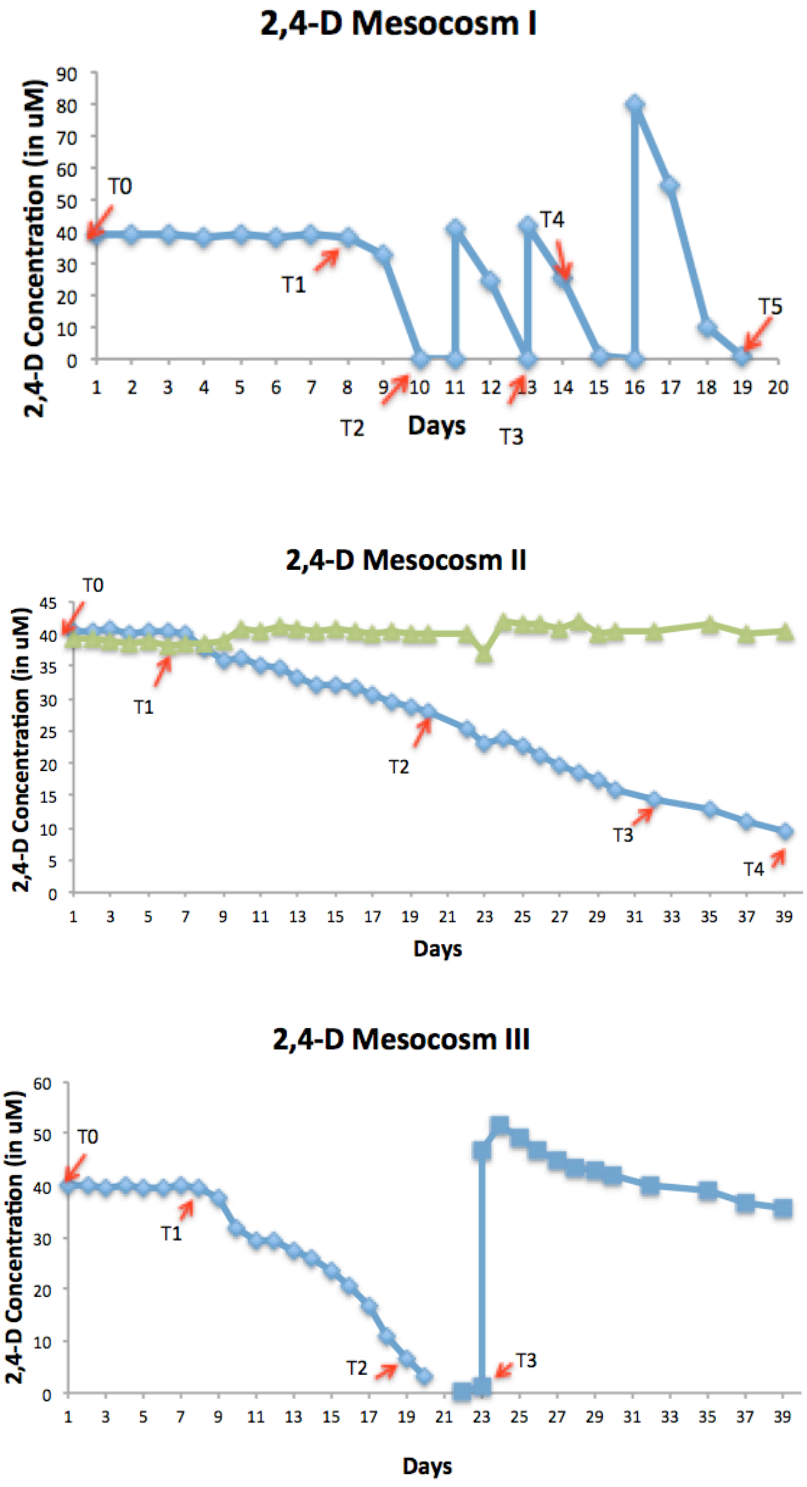

Figure 1. **Degradation profiles of the triplicate 2,4-D mesocosms.** The blue line in each figure indicates the 2,4-D concentration over time, and the green line represents the negative control (2,4-D in sterilized lake water). The red arrows in each figure indicate the time points at which DNA was extracted. The population at the last time point for

each mesocosm was used as the inoculum for isolation work (i.e., E1-T5, E2-T4, and E3-T3).

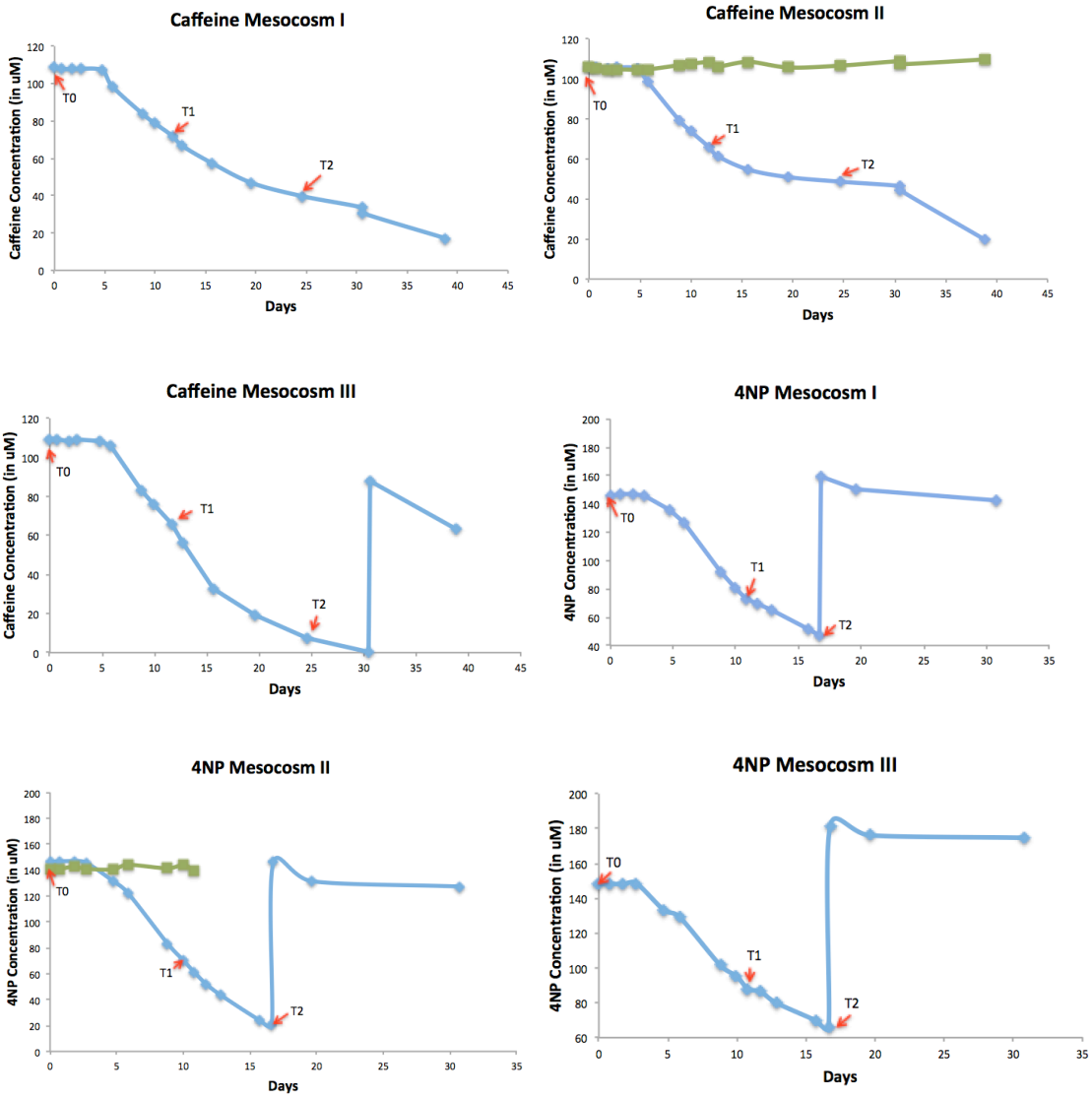

Figure 2. **Degradation profiles of caffeine and 4-nitrophenol (4-NP) mesocosms.** The blue line in each figure indicates caffeine and 4-NP concentration over time, and the green line represents the negative control (caffeine and 4-NP in sterilized lake water). The chemical concentration of added compounds in each mesocosm experiments was measured by high-performance liquid chromatography (HPLC), and the red arrow in each figure indicates the time points at which DNA was extracted. For 4-NP, the DNA at T = 2 in mesocosm I was sequenced. For caffeine, the DNAs at T=2 in mesocosm I and III were sequenced. The caffeine degraders were isolated from mesocosm 1 at T = 2 (Caff-E1-T2), and 4-NP degraders were isolated from meosocosm 1 at T = 2 (4NP-E1-T2).

High-performance liquid chromatography (HPLC) analysis

The concentration of 2,4-D, 4-NP and caffeine in the mesocosms were measured by HPLC. All mesocosms were well mixed on stir plates for approximately 30 minutes before ~1-mL samples were taken. Samples were centrifuged at 14,000 rpm for 5 minutes, and the ~400-µL of the supernatant was removed for further analysis. HPLC was performed on an Agilent 1100 system (Santa Clara, CA, USA) equipped with a diode array detector, an autosampler, and an Ascentis Express C18 (Reserved Phase) HPLC column (Bellevue, WA, USA). The autosampler and column heaters were maintained at 4°C and 45°C, respectively. The HPLC analytical protocols were as follows. For 2,4-D, the mobile phase consisted of a 50:50 ratio of 0.5% trifluoroacetic acid in water to 0.05% trifluoroacetic acid in acetonitrile. For caffeine, the mobile phase consisted of a 65:35 ratio of 0.1% trifluoroacetic acid in water to 0.05% trifluoroacetic acid in acetonitrile. For 4-NP, the mobile phase consisted of a 90:10 ratio of 0.1% trifluoroacetic acid in water to 0.05% trifluoroacetic acid in acetonitrile. For all methods, the autosampler and column heaters were maintained at 4°C and 45°C, respectively.

Isolation and growth of bacteria

An 1 L sample was saved from the last sampling time point from all 2,4-D mesocosms (E1-T5, E2-T4, and E3-T3), 4-NP (E1-T2), and caffeine (E1-T2) for isolation procedures. The 2,4-D degraders were isolated using 10-fold serial dilutions in 12-well plates filled with selective enrichment medium, i.e., liquid ¼-strength minimal salts medium (MSB) (Stanier 1942), containing 7.6 mM $(NH_4)_2SO_4$ as nitrogen source with 40µM 2,4-D. The cultures were incubated at room temperature without shaking and covered with aluminum foil. A 100 µL culture volume from the most diluted well that

showed a turbidity increase was spread to MSB agar (1.5%) plates containing 1 mM 2,4-D (the higher concentration was used to obtain more robust 2,4-D degraders), and incubated at room temperature. Individual colonies that appeared after 5 days of incubation were tested for the ability to degrade 2,4-D in liquid MSB with and without carbon or nitrogen source. The confirmed colonies were re-streaked consecutively at least four times to ensure purity. Isolates that used 2,4-D as the sole source of carbon and energy were selected for further study. Similarly, the caffeine and 4-NP degraders were isolated by 10-fold serial dilutions in 96-well plates. After 7 days, the disappearance of chemicals was analyzed visually for 4-NP (color of media changing from yellow to transparent) and by HPLC for caffeine. A couple ml of the most diluted well that showed disappearance were spread onto ¼-strength nitrogen-free MSB (Stanier, Palleroni et al. 1966) agar plates (1.5%) containing caffeine (100 μM) or 4-NP (150 μM); caffeine or 4-NP served also as nitrogen source. All caffeine and 4-NP degraders were confirmed on carbon- and nitrogen-free MSB agar plates (1.5%). All isolates were preserved at -80°C with ~20% glycerol.

Isolate identification

The 16S ribosomal RNA genes from selected isolates were amplified by PCR using a previously designed universal 16S primer set (8F forward primer 5'-AGAGTTTGATCCTGGCTCAG-3' and 1492R reverse primer 5'-GGTTACCTTGTTACGACTT-3') (Turner, Pryer et al. 1999) and the thermocycling profile as follows: 98°C for 3 minutes, 30 cycles of 98°C for 10 seconds, 55°C for 45 seconds, 72°C for 2 minutes and 10 seconds, and 72°C for 10 minutes, using The

amplification products were sequenced by Sanger sequencing (Genewiz, Plainfield, NJ, USA). The raw sequences were trimmed using SolexaQA (Cox, Peterson et al. 2010) (v3.1.3) (-h 20) with default settings, and merged -if overlapping- using PEAR (Zhang, Kobert et al. 2014) (v0.9.6), with default settings and p value of 0.0001. The NCBI database and BlastN search (Altschul, Gish et al. 1990) were used to identify the best match of the resulting sequences.

Genomic DNA extraction and sequencing of 2,4-D degrading isolates

The 2,4-D degraders were grown on ¼-strength liquid MSB containing 1.4 mM 2,4-D for about 60 hours at room temperature on a shaker (i.e., ~150 rpm). Three mL of the culture (OD, ~0.2) was spun down at 7500 rpm for 5–10 minutes, and the DNA pellet was washed twice with 1× phosphate-buffered saline (PBS). The QIAamp DNA Mini Kit (Qiagen, Valencia, CA) was used for genomic DNA extraction as described by the manufacturer. DNA was eluted in 100 µL of buffer AE (10mM Tris, 1mM EDTA pH=8.0). Genomic DNA was extracted from 4-NP and caffeine degraders using the same kit. Isolate genomes were sequenced as described for 4-NP and caffeine mesocosm samples above.

Sequence analysis and metagenomic assembly

An in-house trim pipeline (trim.pbs) was used for quality check and trimming (available through http://enve-omics.gatech.edu). Briefly, the genomic and metagenomic sequences of isolates were trimmed using SolexaQA++ (Benson, Karsch-Mizrachi et al. 2004) with a PHRED score cutoff (-h) of 20 and a minimum fragment length of 50 bp

(Hyatt, Chen et al. 2010). Only coupled reads with both sisters longer than 50 bp after trimming were used further. The statistics of the trimmed datasets and their assemblies for each metagenome and genome are provided in Supplementary Table S2 and S6, respectively. The average coverage of each metagenomic dataset was estimated using Nonpareil with default parameters (Supplementary Figure S2) (Rodriguez-R and Konstantinidis 2014). Genomic and metagenomic reads were assembled following a previously described hybrid-assembly protocol (Luo, Tsementzi et al. 2012). Briefly, short sequences were assembled with Velvet (Zerbino and Birney 2008) and SOAP (Luo, Liu et al. 2012) de novo using k-mer values ranging from 21 to 63. The three assemblies from each algorithm that maximized the N50 and/or the fraction of assembled reads were chosen for subsequence assembly with Newbler (v2.0), which merged (co-assembled) the resulting assemblies.

16S rRNA gene sequence extraction and metagenome composition analysis

Metagenomic reads encoding 16S rRNA gene (16S) fragments were extracted using Parallel-Meta (version: v2.4) (Su, Pan et al. 2014) for all time points with default settings, except that the RDP 16S rRNA gene database was used as reference (option -d R). The 16S-encoding sequences were identified taxonomically using the QIIME pipeline (version: v1.8.0) (Caporaso, Kuczynski et al. 2010), with full-length closed-reference operational taxonomic unit (OTU) picking at 97% nucleotide sequence identity based on the August 2013 GreenGenes release 13-5 database (DeSantis, Hugenholtz et al. 2006). The pick_closed_reference_otus.py script was used with the following parameters: enable_rev_strand_match = True, max_accepts = 1, max_rejects = 8, stepwords = 8, and

word_length = 8. A representative sequence from each OTU was aligned against the GreenGenes reference sequences using PyNAST with default settings (DeSantis, Hugenholtz et al. 2006), after removing gaps in the alignment. Sequences that failed to align to reference taxa using the RDP (-m rdp) classifier at 50% confidence (Wang, Garrity et al. 2007) were also removed from further analysis. For alpha and beta diversity estimates, a subset of sequences obtained by randomly subsampling each dataset at the same depth (that of the smallest dataset: for 2,4-D, E3-T3 with 6,189 sequences, and for caffeine and 4-NP: Caff-E1 with 1,301 sequences) was used. Alpha diversity was measured by three metrics, i.e., observed OTUs, Shannon diversity, and whole-tree phylogenetic diversity, as implemented in QIIME with default settings (Supplementary Table S3). Beta diversity was measured by binary Sorensen-Dice metrics (Supplementary Figure S3).

Scaffolding of *Burkholderia* sp. KK1 plasmid contigs

Two previously described 2,4-D degrading plasmid nucleotide sequences were downloaded from GenBank, pJp4 (accession no. AY365053.1) (Trefault, De la Iglesia et al. 2004) and pM7012 (accession no. NC_022995.1) (Sakai, Ogawa et al. 2014), and used to scaffold the contigs from the *Burkholderia* sp. KK1 isolate genome assembly. KK1 plasmid contigs scaffolding was performed manually, contig by contig, according to BlastN-based matched coordinates on reference plasmid sequences, and all connections between contigs were confirmed by pair-end sequencing read mapping (requiring a >40 bp alignment of 100% nucleotide identity cutoff for a matching read). Several contigs were not included in the final scaffold due to ambiguous read mapping caused by the

presence of highly repetitive sequences, transposases, and integrases (Supplementary Table S4). The *tfd* gene cluster comparison among the three plasmids, the putative KK1 plasmid, plasmid pJp4, and plasmid pM7012, was performed using Easyfig (v.2.1) (Sullivan, Petty et al. 2011) (see Supplementary Figure S4). The resulting scaffolded plasmid sequence was further validated using the plasmid constellation network (PLACNET, v1.01) (Lanza, de Toro et al. 2014) with default output settings (Supplementary Figure S7). The input files for PLACNET were assembled contig sequences, length and coverage files and genome-mapping file using BWA-MEM algorithm (Li and Durbin 2009).

Annotation of genomic and metagenomic sequences

Prodigal (v2.6.2) (Hyatt, Chen et al. 2010) with the single model (-p single) was used to predict proteins in isolate genomes. The Blast2Go (Conesa, Götz et al. 2005) was used to annotate genomic proteins using standard pipeline (Blast, Mapping, and Annotation). A phylogenomic approach was used to further corroborate the Blast2Go findings and identify 2,4-D, 4-NP, and caffeine degradation genes in the isolate genomes as follows. Well-characterized reference sequences of genes that encode 2,4-D, 4-NP, and caffeine catabolic enzymes were downloaded from GenBank (Benson, Karsch-Mizrachi et al. 2004). A BlastP search was then performed of these reference protein sequences against the predicted genome proteins, using default settings with an e-value cutoff of 0.001 and an identity threshold of $\geq$ 35%. All matching sequences were further evaluated by visually inspecting phylogenetic trees built with these sequences and reference protein sequences from GenBank, using the neighbor-joining method.

Matching sequences representing paraphyletic, long branches compared to the reference protein sequence were excluded from the analysis.

For metagenomic datasets, protein-coding sequences were predicted for large contigs (≥500 bp) using Prodigal (v2.6.2) with the Meta model (-p meta). Gene functional annotation was performed by a BlastP search of the predicted protein sequences against the SwissProt database (Wu, Apweiler et al. 2006), and the SEED database using subsystems categories (Overbeek, Begley et al. 2005). Only best matches of at least bit score 60 were considered for functional annotation. MyTaxa (Luo, Rodriguez-R et al. 2014) was used for taxonomic assessment of all large contigs with default parameters.

Assess abundance of isolates and biodegradation genes

The abundances of isolates and genes were assessed by BlastN searches (-xdrop_gap 150, -e-value 0.001) against their corresponding metagenomes using at least ≥150 bp alignment length and ≥99% identity cut-off for a match. The relative abundances of isolates and genes were calculated as the percentage of reads mapped on them in each metagenome. In particular, the length of all read mapping on a reference isolate genome or gene sequence were summed and divided by the total length of the reference sequence. To calculate the genome equivalent for each gene/function, i.e., the fraction of total cells encoding the gene, an in-house ruby script was used, essentially as described previously (Rodriguez-R. L-M, Overholt W. A. et al. 2015). Briefly, the script is used to detect and extract reads encoding any of 101 essential protein-coding genes (universally conserved single-copy) (Dupont, Rusch et al. 2012) from the Genome Property database (entry ID:

GenProp0799, named "bacterial core gene set, exactly 1 per genome") (Haft, Selengut et al. 2005). The ten models, *rpoC*, *pheT*, *proS*, *glyS*, *era*, and *tRNA* synthase class I were eliminated because more than one model represented the same gene family or represented extremely conserved families in terms of sequence conservation. For the remaining 91 essential models, their median sequence depth was determined based on the number of reads mapped to each gene (filter with length ≥80 bp and identity ≥97%) divided by the length of the gene (reads/bp). The genome equivalent of a target gene was estimated as the ratio of its sequence depth, divided by the median sequence depth of the 91 marker genes.

Functional gene profiles

The abundance of each predicted gene/protein sequence in each metagenome was estimated by the number of reads mapped to the gene, normalized for the length of each gene. Read mapping was performed using BlastN (-xdrop_gap 150, -e-value 0.001) and a minimum cut-off for a match of ≥80 bp alignment length and ≥97% nucleotide identity. Functional groups based on SEED Subsystems were identified using the DESeq2 package (Love, Huber et al. 2014). The coverage (X) of all predicated genes in each post-enrichment metagenome (i.e., E1-T5, 4NP-E1, Caff-E3, etc.,) was determined by utilizing an in-house Perl script (i.e., BlastTab.seqdepth.pl which available through http://enve-omics.gatech.edu), and compared to the coverage in the original lake metagenome at T=0. The most enriched genes (fold-change) were clustered based on GO terms.

tfdA primer design, plasmid cloning, and quantification of _tfdA_ and 16S rRNA genes by quantitative PCR (qPCR)

Two *tfdA* genes were identified in the *Burkholderia* sp. KK1 assembled plasmid, and a primer set was designed based on the alignment of the two *tfdA* genes. (Supplementary Table S5). One *tfdA* gene, which showed 100% nucleotide identity to a homolog in the reference pM7012 plasmid (Sakai, Ogawa et al. 2014), had a 100% identity match to both forward and reverse primers. The other *tfdA* gene, which had a 100% identity to a homolog in the reference pJp4 plasmid (Trefault, De la Iglesia et al. 2004), had a 100% identity to the forward primer, but had a 2-bp mismatch at the 5′ end of the reverse primer; therefore, the primer set was expected to amplify both genes but not necessarily with the same efficiencies. Another version of *tfdA* gene was identified from mesocosm I metagenomic contig01889 (denoted as contig01889_tfdA; Supplementary Table S5); however, the alignment between the primer set and this metagenomic *tfdA* gene was low, especially for the reverse primer (95% nucleotide identity with forward primer and 42% identity with reverse primer).

The TOPO-TA cloning kit (pCR2.1-TOPO vector; Invitrogen, Carlsbad, CA, USA) was used to make the qPCR standard plasmid and the single-copy *tfdA* (142 bp) insertion was confirmed by Sanger sequencing. The *tfdA* cloning was performed following the methods described below. Briefly, the 142 bp fragment of the *tfdA* gene amplicon was cloned into the pCR2.1 vector as directed by the manufacturer and subsequently introduced into *Escherichia coli* DH5α cells by heat-shock at 42°C for 30 seconds (following the instruction manual, Thermo Fisher Scientific). The transformed

cells were grown in 250 µL of SOC medium at 37°C for 1 hour with shaking (~220 rpm), and 50 µL of the transformation mixture was spread on a fresh LB plate containing 100 µg/mL ampicillin and incubated overnight. Four single colonies were streaked to a fresh LB plate containing ampicillin (100 µg/mL), and two independent isolates were chosen for growth (5 mL of LB containing 100 µg/mL ampicillin). Plasmid extraction was performed after 12 hours of growth at room temperature on a shaker (~200 rpm) with 2 mL of the overnight culture using the QIAprep Miniprep Kit, and cloning was confirmed by Sanger sequencing.

For qPCR, *tfdA* primer concentrations were optimized; *tfdA* primer concentrations of 100, 250, and 500 nM were tested simultaneously using the following qPCR thermocycling profile: 50°C for 2 minutes, 95°C for 10 minutes, and 40 cycles of 95°C for 15 seconds and 60°C for 1 minute. The components of the reaction mixture were as follows: 2 µl of template DNA, forward and reverse primer at stated concentrations, 1X SYBR Green PCR Master Mix (ThermoFisher Scientific) and water (Total volume = 20-µL). The 250 nM primer concentration yielded a robust standard curve (R^2 = 0.999 and efficiency >98%), and thus, was used for subsequent work. Ten-fold serial dilutions of the standard plasmid were used to generate a standard curve over the dynamic range from 44 gene copies to 4.4E07 gene copies, respectively. The qPCR quantification of *tfdA* was first tested and validated with a serial dilution experiment of *Burkholderia* sp. KK1 genomic DNA before being applied to mesocosm time-series DNA samples (T = 0, 1, 2, 3, 4, and 5). To validate the qPCR assay, the *Burkholderia* sp. KK1 abundance estimate obtained by qPCR (1.96E08 cells/mL) was compared with the expected abundance based

on the dilution factor of the sample and microscope-based direct cell counting with DAPI (1.19E08 cells/mL). All unknown samples, standards, and a blank were assayed in triplicate. The average of the triplicates was used to quantify the abundance of *tfdA* genes obtained by qPCR (converted to number of *tfdA* copies per ml). The environmental DNA samples were diluted in water with a 32-fold dilution factor (DF), since the undiluted samples inhibited the qPCR amplification.

Total 16S rRNA genes were also quantified by qPCR to examine shifts in the total number of bacteria during perturbations (assuming a stable average 16S rRNA gene copy number) in 2,4-D mesocosm I. The standard plasmid used was the pBva1-16S. The primer set was designed previously (Ritalahti, Amos et al. 2006) (1055yF and 1392R, Supplementary Table S5). The primer concentration was 1000 nM and same qPCR cycle was used as for the *tfdA* gene above.

CHAPTER 3

RESULTS

Mesocosm biodegradation profiles

The three replicate mesocosms revealed three distinct 2,4-D biodegradation profiles. In Mesocosm I, 2,4-D degraded within ~10 days. After three 2,4-D re-spikes, degradation was robust and faster (~2 days), even after a doubling in the spiked-in 2,4-D concentration during the last re-spike (Figure 1). In contrast, mesocosms II and III showed incomplete 2,4-D biodegradation. Mesocosm II never completely degraded 2,4-D, even after a 40-day period, while mesocosm III demonstrated complete degradation within 21 days, and weak degradation after one re-spike. Variability was also observed in the 4-NP and caffeine mesocosms, albeit degradation profiles were more similar among the three replicate mesocosms in these cases compared to the 2,4-D mesocosms (Figure 2). Degradation of 4-NP and caffeine took a much longer period, at least 20 days, and was often incomplete, while re-spike events typically showed even less robust degradation.

To test whether the observed differences in the 2,4-D degradation profiles were attributable to nutrient limitation (e.g., nitrogen) or change of pH (e.g., the final pH in mesocosm II and III were below 6), small volumes (~2 mL) from mesocosms II and III were removed and supplemented with ¼-strength MSB minimal media containing 7.6 mM $(NH_4)_2SO_4$. Both resulting microcosms showed complete 2,4-D removal in less than a day. Mesocosm I apparently did not require the addition of nutrients. The reason

underlying the variation observed in degradation profiles of 2,4-D, 4-NP and caffeine was presumably due to lack of nutrients (e.g., nitrogen) and/or lower pH, which apparently affected the response of rare biosphere, e.g., different rare taxa or genes were activated in each mesocosm.

Degraders are members of the rare biosphere

To determine whether degradation of the added organic compounds was due to abundant or low-abundance member(s) of the microbial community, 2,4-D degraders were isolated from the three mesocosms at the last sampling point (E1-T5, E2-T4, and E3-T3). Their genomes were sequenced and were assignable to *Burkholderia*, *Sphingopyxis*, or *Variovorax* genera for mesocosm I, II and III, respectively, based on 16S rRNA gene sequence or whole-genome-based ANI (Goris, Konstantinidis et al. 2007) analysis. The relative abundances of the degraders during the mesocosm incubation were assessed based on the number of metagenomic reads from each sampling point mapping onto their genome sequence at high stringency (>99% nucleotide identity, excluding the rRNA gene operon, which is highly conserved). All three 2,4-D degraders were under detection limit at the T = 0 metagenome of their corresponding mesocosm, i.e., 1.372×10^{-4}, p value=0.002, based on a recently developed algorithm by our team (Castro et al., in review). However, upon 2,4-D addition, all three 2,4-D degraders, especially the *Burkholderia* sp. KK1-like population became abundant by at least four orders of magnitude by the time that the added 2,4-D was completely biodegraded (i.e., making up ~0.24% of total community at T=2). Abundance was even higher at the last sampling time point (T = 5), i.e., after three 2,4-D re-spikes, making up ~0.82% of total

community. Similar results were observed for 4-NP and caffeine degraders (see below). This level of enrichment was also consistent with theoretical calculations based on the energy available for growth on 2,4-D as a sole source of carbon and energy (see supplementary information). These results confirmed our hypotheses that the added compounds selected for rare community members in original lake inoculum (Figure 3).

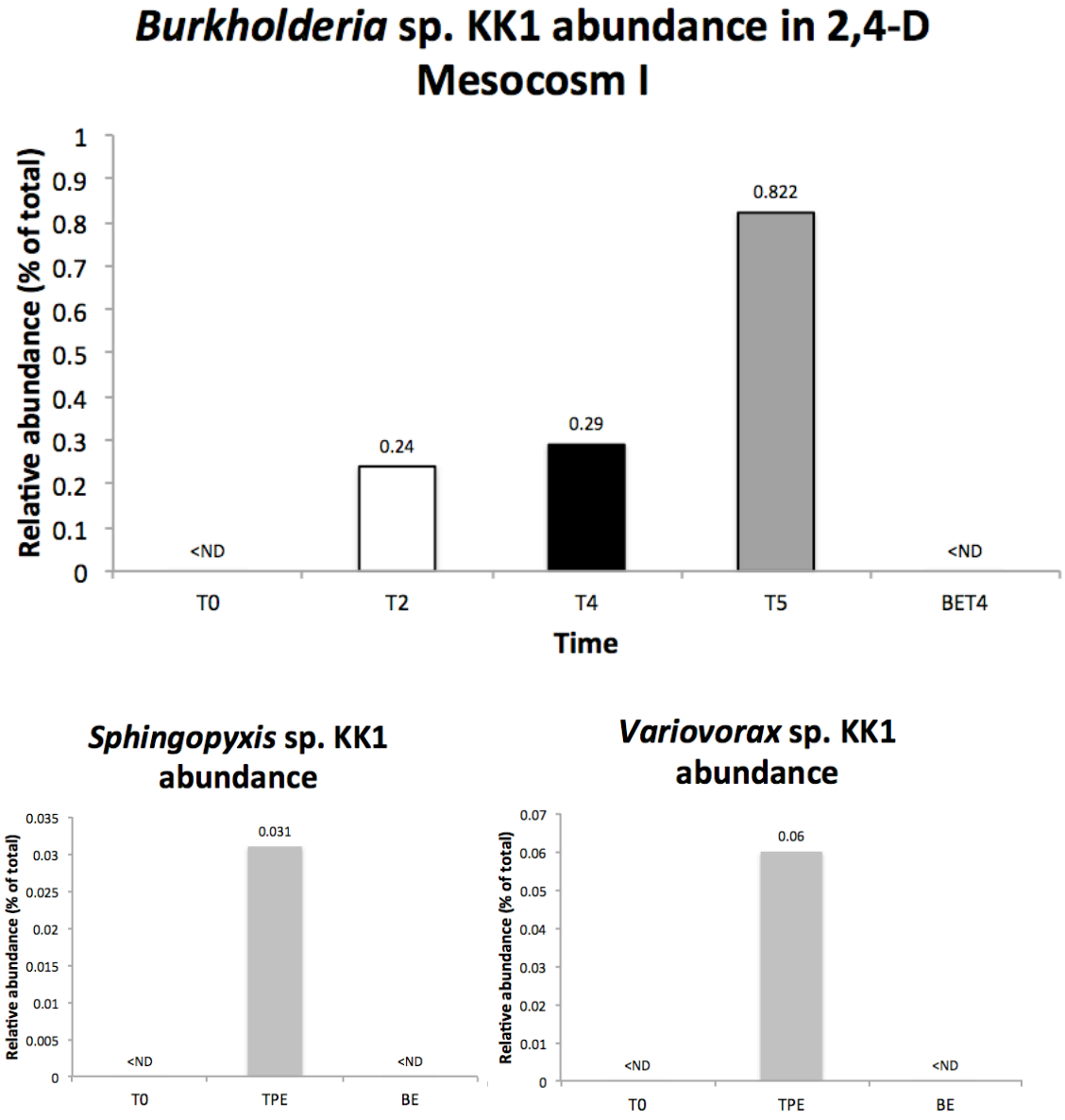

Figure 3. **Relative abundance of 2,4-D degraders from each mesocosm**. The abundance was calculated as the fraction of total metagenomic reads mapping on the corresponding genome sequences (identity ≥99% and length ≥80 bp) in each metagenomic dataset. BE stands for Bottle Effect, and TPE stands for the last sampling time point, from which the degraders were also isolated.

The 4-NP degraders were isolated from 4-NP mesocosm I (4NP-E1) at T = 2 and were assigned to *Pseudomonas*. The predominant caffeine degrader were isolated from caffeine mesocosm I (Caff-E1) at T = 2 and assigned to *Pelomonas*. During enrichment, these degraders became abundant, similar to the results reported above for 2,4-D degraders, albeit the level of enrichment of the caffeine degrader was an order of magnitude less pronounced (Supplementary Figure S5).

Abundance and phylogeny of 2,4-D, 4-NP and caffeine biodegradation genes

The known genes responsible for 2,4-D degradation are the *tfdABCEDF* gene cluster (Top, Holben et al. 1995), and the *cadABCD* gene cluster (Nielsen, Xu et al. 2013). The *tfd* gene cluster was present in the genome of the *Burkholderia* sp. KK1 isolate from mesocosm I and in assembled contigs from all mesocosms after perturbation. The *cad* gene cluster was present in the *Sphingopyxis* sp. isolate, and was detected in lower abundance in metagenomes from mesocom I at T=5 relative to those from in mesocosms II and III. Three *tfd* operons were identified in the genome of *Burkholderia* sp. KK1. Two of them, one complete and one that contained only *tfdBCE* genes, showed ~100% nucleotide identity to those in the previously reported pM7012 plasmid (Sakai, Ogawa et al. 2014). The third complete *tfd* gene cluster in KK1 genome showed ~100% nucleotide identity to that of those in the previously reported pJp4 plasmid (Trefault, De la Iglesia et al. 2004).

The abundance of 2,4-D biodegradation genes during enrichment was assessed by querying the sequences of each allele (variety) recovered in the genomes or assembled

contigs against the time-series metagenomes from the mesocosms. Results for *tfdA* genes are preferentially reported below because it serves as biomarker of 2,4-D biodegradation, which converts 2,4-D to 2,4-dichlorophenol (2,4-DCP), known as the first step of 2,4-D biodegradation. Consistent with the whole-genome-based abundance of the isolates reported above, no *tfd* gene was detectable in the original inoculum (time=0 metagenome), while the percentage of total cells (i.e., genome equivalents) encoding 2,4-D biodegradation genes increased over time (Figure 4 for 2,4-D; and Supplementary Figure S8 for 4NP and caffeine catabolic genes). Phylogenetic analysis of the *tfdA* genes recovered in assembled contigs from all three 2,4-D mesocosms and isolates revealed however that, while the two alleles of *tfdA* present in KK1 genome were present in the metagenomes, the most abundant *tfdA* alleles in the metagenome formed a new cluster that was divergent from previously identified *tfdA* genes or those present in KK1 (i.e., 83% and 92% amino acid identity identity, respectively) (Figure 5; underlying alignment is shown in Supplementary Table S5).

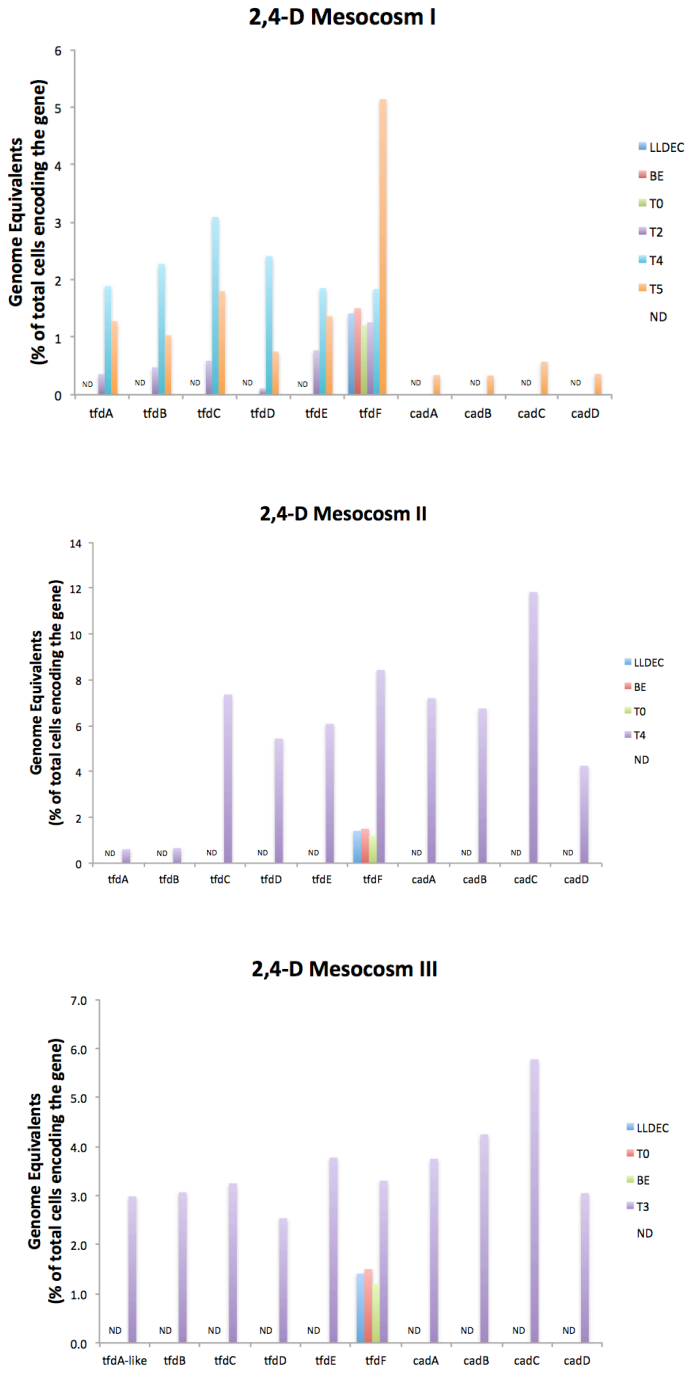

Figure 4. **Abundance of 2,4-D biodegradation genes during mesocosm incubation**. Genes (x-axis) forming the *tfd* and *cad* operons involved in 2,4-D degradation. The abundance was measured by genome equivalents (% of total cells encoding the gene) at each time point. ND: Not Detected.

The phylogenetic analysis also revealed several additional "*tfdA*-like" genes present in the 2,4-D mesocosms, which are distant homologs (e.g. showing <40% amino acid identity, Figure 5) of experimentally verified *tfdA* genes. These *tfdA*-like genes (also known as *tfdAα*) (Hogan, Buckley et al. 1997) are ubiquitously found in both 2,4-D degraders and non-2,4-D degraders, and the proteins that the genes encode, TfdAα, likely do not contribute to 2,4-D biodegradation. For instance, it has been reported that TfdAα proteins show significantly weak α-ketoglutarate-dependent 2,4-D dioxygenase activity compared to TfdA protein of pJp4 (1/1,000 of specific activity) (Itoh, Kanda et al. 2002). Thus, these gene alleles were not discussed further although several of them increased in abundance during our incubation, indicating that they might be involved in 2,4-D biodegradation.

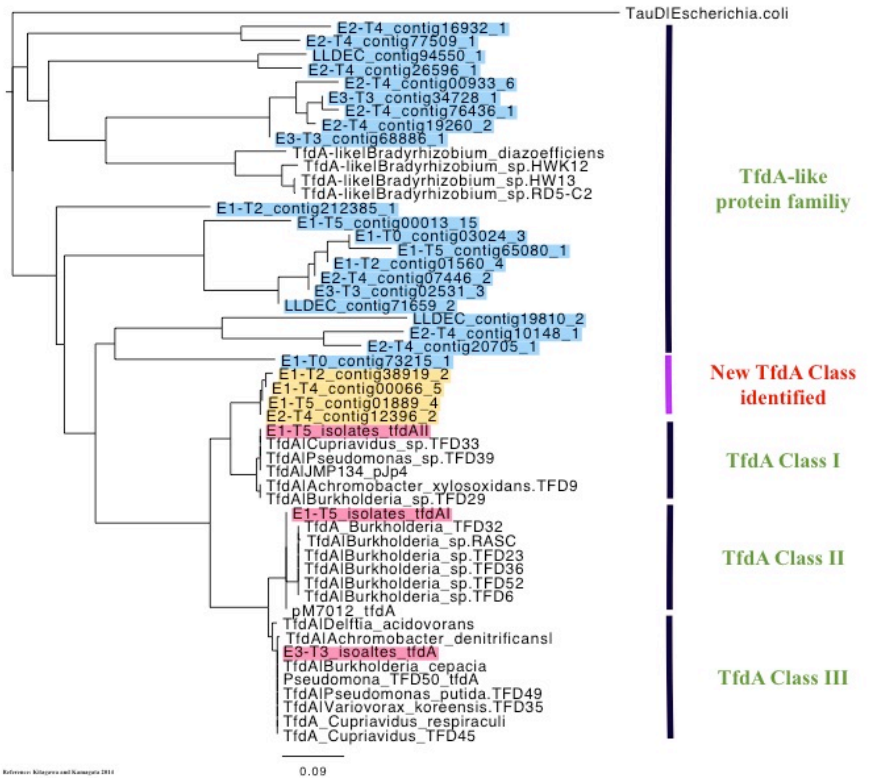

Figure 5. **Phylogeny of all *tfdA* and *tfdA*-like genes recovered from the three 2,4-D mesocosms and isolates**. The tree was built using the neighbor-joining method as implemented in Geneious (v.8.1.8) based on an amino acid alignment. Blue denotes all *tfdA*-like genes recovered from the three mesocosms. Yellow denotes *tfdA* genes that form a new cluster compared to previously described *tfdA*. Pink denotes *tfdA* genes encoded by our isolates.

In addition to *tfd* genes, the *cad* gene cluster was also enriched during selective growth in all 2,4-D mesocosms, especially in mesocosms II and III (Figure 4). The enzymes encoded by this complex are thought to catalyze the initial step of 2,4-D biodegradation in *Bradyrhizobium* sp. strain HW13 (analogous to the function of *tfdA*) (Kitagawa, Takami et al. 2002). *Cad* gene clusters were only identified in the *Sphingopyxis* sp. isolate obtained form mesocosm II (i.e., *cadABCD* genes on genomic

contig00006). Given also the abundant *tfdA* genes based on metagenomes that were not encoded by any of the isolates, these results revealed that several additional 2,4-D degraders were present in mesocosms I and III, but evaded cultivation under the laboratory conditions tested. In summary, our results suggested that various alleles of 2,4-D biodegradation genes (*tfd* or *cad*) and distinct populations were enriched in the three mesocosms and that substantial gene functional redundancy maybe present among the low-abundance organisms, even in relative small volumes of lake water.

4-NP can be degraded by either the hydroquinone (HQ) (Spain and Gibson 1991) or the hydroxyquinol (BT) pathways (Zhang, Sun et al. 2012). The HQ pathway is the predominant pathway in gram-negative bacteria, such as *Pseudomonas* sp. strain WBC-3 (Liu, Zhang et al. 2005). The HQ genes cluster *pdcABCDEFG* was found in two pieces, on contig26 (*pdcABDE*) and contig116 (*pdcCDEFG*) in the 4-NP degrading *Pseudomonas* sp. isolate, showing ≥ 90% nucleotide identity to the previously described gene cluster. Several *pdc* genes were also identified in assembled metagenomic contigs from the 4-NP mesocosm I metagenome (e.g., *pdcA* genes encoding protein from contig00632 with 89.28% amino acid identity to the previously described proteins), and the abundance of all *pdc* genes increased from undetectable levels to about 1% of the total cells encoding the genes during the incubation time, with the exception of *pdcF* (Supplementary Figure S8). Notably, the coverage of all 4-NP genes (i.e., 5.39X) was substantially higher than that of the isolate genome (i.e., 3.11X), indicating that additional 4-NP degraders that were not isolated were present in the mesocosms.

Caffeine degradation occurs via the *N*-demethylation pathway, which contains five enzymes *ndmABCDE* (Summers, Louie et al. 2012). The *ndm* genes were previously described in the caffeine degrader *Pseudomonas putida* CBB5, encoded on a 13.2-kb genomic DNA fragment (Summers, Louie et al. 2012, Summers, Seffernick et al. 2013). Dash and Gummadi also reported that the *N*-demethylation enzymes (e.g., *ndmABCDE*) are encoded on a 12-kb plasmid, but the sequences are not available (Dash and Gummadi 2006). Homologs of the *ndmABCDE* genes were identified on contig10 in the *Pelomonas* sp. isolate genome with ≥ 50% amino acid identity to the previously reported *ndmABCDE* (Summers, Louie et al. 2012). Caffeine degradation can also occur via C-8 oxidation pathway (Mohanty, Yu et al. 2012). One of the key enzymes for the latter pathway is caffeine dehydrogenase (*cdh*), which contains three subunits (i.e., *cdhABC*) (Yu, Kale et al. 2008). The abundance of *ndmABCDE* and *cdhABC* genes also increased during mesocosm incubation, albeit the level of increase was not consistent among all genes of the *ndmABCDE* operon, unlike the 4-NP and 2,4-D degradation genes (Supplementary Figure S8).

qPCR analysis of 2,4-D degradation genes

To further validate the metagenome-based results, *tfdA* gene sequences present in the genome of *Burkholderia* sp. KK1 isolate were quantified by qPCR of the mesocosm community DNA samples. In general, the results were consistent with the metagenomic findings. In particular, at two early time points of the 2,4-D mesocosm I (T = 0 and T = 1), the abundance of *tfdA* genes was not quantifiable (i.e., below the detection limit of 44

gene copies), especially for T=0 sample. At later time points (T = 4 and T = 5), the two *tfdA* genes increased in abundance by at least one order of magnitude (when compared to the detection limit of 44 copies of the *tfdA* gene) and were robustly detected by qPCR (Figure 6). These results revealed that the *tfdA* genes were present in a rare fraction of the original lake inoculum, and became abundant after perturbation.

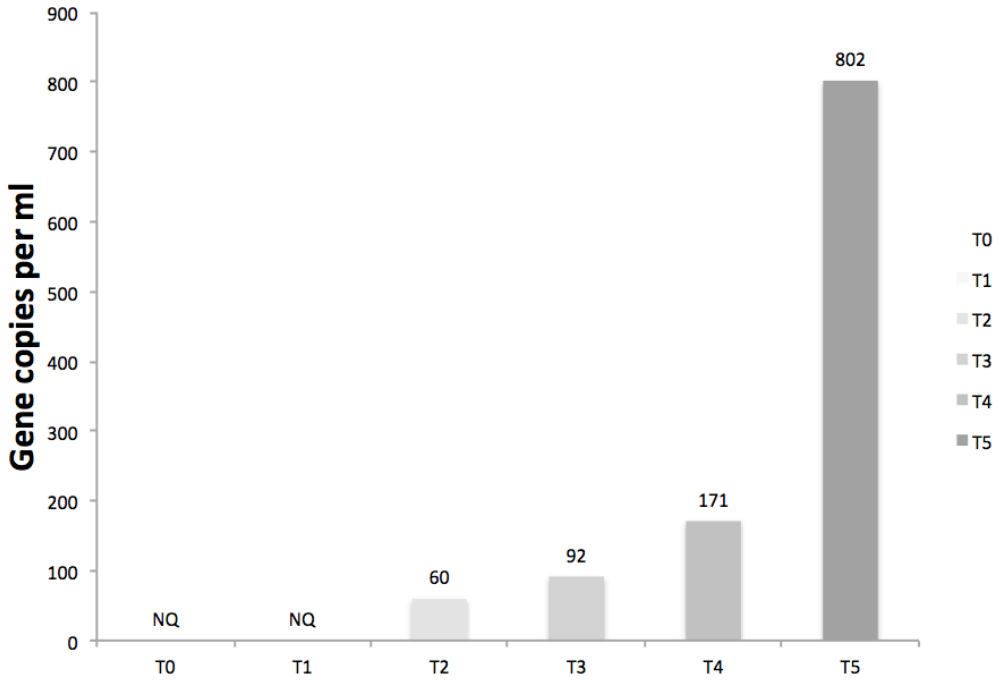

Figure 6. **Quantification of *Burkholderia* sp. KK1 *tfdA* genes in mesocosm I based on qPCR.** qPCR analysis was performed on samples from 2,4-D mesocosm I and abundance was estimated against a standard curve of a plasmid encoding *tfdA* as described in the Material and Methods section. NQ: PCR product was obtained but was not quantifiable; see text for details.

Genetic basis of *Burkholderia* sp. KK1 2,4-D biodegradation genes

Scaffolding analysis of the *Burkholderia* sp. KK1 contigs against available sequences (Sakai, Ogawa et al. 2014) (Trefault, De la Iglesia et al. 2004) suggested that all *tfd* genes were present on a putative megaplasmid (~580 kb). Plasmid genes encoding

a Walker-type ATPase (Escobar‐Paramo, Faivre et al. 2009) and a centromere-binding protein (*parB*) were also found in contigs, which linked to the contigs that encoding the *tfd* gene cluster. Considering that the structure of the KK1 plasmid showed high similarity with pM7012 and pJp4 transmissible plasmids from *Burkholderia* sp. M701 (Sakai, Ogawa et al. 2014) (two-way AAI=99.25% SD=6.25%) and *Ralstonia eutropha* JMP134 (Trefault, De la Iglesia et al. 2004) (one-way AAI=47.97% SD=15.60%) (Supplementary Figure S4) and shared all *tra* genes required for transmissibility except for *traI* (relaxase), the KK1 plasmid is also probably mobile (or was mobile in the recent past). Further, the ~100-kb long region, which included the three *tfd* gene cluster, (highlighted regions in Supplementary Table S4) also encoded many direct and inverted repeated sequences, insertion sequences, transposases, phage integrases, and many hypothetical proteins. Thus, it is appears as if this region represents a highly dynamic and mosaic part of the genome.

Furthermore, we also obtained a *Novosphingobium* sp. (Family: *Alphaproteobacteria*) isolate able to degrade 2,4-D from our enrichments. Its genome sequence revealed *cad* genes showing 100% nucleotide identity with a homolog from a previously described conjugative plasmid pCADAB1 encoded by the *Sphingobium* sp. ERG5 isolate (*Alphaproteobacteria*) (Supplementary Figure S6), (Nielsen, Xu et al. 2013). The *cad* gene cluster identified in the *Novosphingobium* sp. isolate was also flanked by mobile gene elements. Collectively, these results indicated that the 2,4-D degradation genes are transferred horizontally, either during the enrichment process in our mesocosms or shortly before that, within the Lake Lanier.

Microbial community taxonomy and functional shifts during enrichment

To assess community-wide responses to the added organic substrates, the taxonomic and functional gene content shifts at the whole-community level were also examined. Based on both MyTaxa assignement of metagenomic contigs (Figure 7) and QIIME analysis of 16S fragments encoded in metagenomic reads (Supplementary Table S3), the community diversity of the 2,4-D mesocosms was much higher at early time points than later time points, presumably due to the strong selection pressure of the substrate and random death due to restricted growth conditions (e.g., lack of primary production). Overall, 515 OTUs were identified in the T=0 metagenome, but at post-enrichment, there were only 339 OTUs remaining, i.e., ~34.2% total OTUs were undetectable from T=0 to T=5. The *Burkholderiaceae* OTU showed among the strongest enrichment in mesocosm I, and the longest 16S rRNA gene sequence fragment (i.e., 264bp) of this OTU showed 100% nucleotide identity to the *Burkholderia* sp. KK1 isolate. Notably, *Comamonadaceae* OTU also increased in abundance relative to the T=0 metagenome. The representative 16S rRNA sequence of this OTU showed >99% nucleotide identity to the previously identified *Delftia acidovorans* strain P4a (Hoffmann, Kleinsteuber et al. 2003), which can completely degrade 2,4-D using its *tfd* gene clusters encoded on a catabolic transposon.

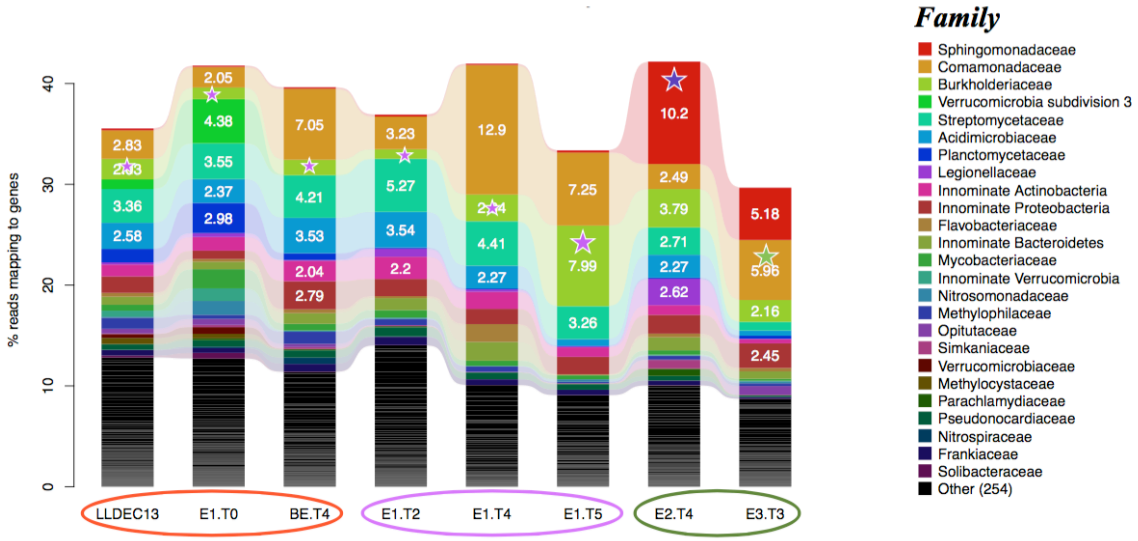

Figure 7. **Microbial community compositions of the 2,4-D mesocosms.** Results shown are based on total 16S rRNA gene-encoding reads recovered from each metagenomic datasets and classified in taxa at the family level. Only taxa that recruited more than >2% of total reads are shown; white number represent relative abundance. The datasets are as follows: original lake (LLDEC13), 2,4-D mesocosm I at T = 0, 2, 4, and 5; 2,4-D mesocosm II at T = 4; 2,4-D mesocosm III at T = 3; Bottle-Effect mesocosm at T = 4. The purple star denotes the family that the *Burkholderia* sp. KK1 isolate was assignable to (100% 16S identity). The blue and the green stars denote the family that the *Sphingopyxis* sp. KK1 and the *Variovorax* sp. KK1 isolates were assignable to (99% and 100% 16S identity, respectively).

There were 445 and 430 OTUs identified in post-enrichment samples from mesocosm II and III, respectively, which indicated a ~13.6% and ~16.5% reduction in OTU richness compared to that at T=0. Considering that the 2,4-D biodegradation was incomplete in these two mesocosms as opposed to mesocosm I, this decrease was consistent with our expectation as it is less significant than the decrease in mesocosm I, which robustly degraded 2,4-D with associated growth of several populations. In mesocosms II and III several taxa that are not known to include 2,4-D degraders were enriched over time (Figure 7). These groups included *Legionellaceae* (*Gammaproteobacteria*), *Comamonadaceae* (i.e., represented by our isolates *Variovorax*

sp.) (*Betaproteobacteria*) and *Sphingomonadaceae*, i.e., represented by our isolates *Sphingopyxis* sp. (*Alphaproteobacteria*),while *Planctomycetaceae* (*Planctomycetes*) and *Acidimicrobiaceae* (*Actinobacteria*) decreased in abundance. It should be noted that *Legionellaceae* encompassed one genus, with about 50 named species, mostly of clinical relevance (Lory 2014); hence, it is likely that this family also contained yet-to-be-described 2,4-D degraders or organisms that benefit from byproducts of 2,4-D degradation. On the other hand, members of *Sphingomonadaceae* and *Variovorax* are known to encode 2,4-D biodegradation capabilities, and were previously detected as a major 2,4-D degraders isolated from 2,4-D heavily treated agricultural fields or soil (Ka, Holben et al. 1994) (Tonso, Matheson et al. 1995, Stibal, Bælum et al. 2012). These two groups were represented by our isolates and their 16S rRNA gene sequences showed >98% nucleotide identity to previously identified degraders. Therefore, it appears that the original lake inoculum contained more than one known and also some poorly-characterized 2,4-D degraders, and all of them benefited from 2,4-D directly or indirectly (i.e., using byproducts).

For the 4-NP and caffeine mesocosms, the trends in total OTUs shifts were similar to those of 2,4-D. Several OTUs encompassing known 4-NP degraders were increased in abundance such as *Pseudomonadaceae* (*Gammaproteobacteria*), represented by our *Pseudomonas* sp. isolates (i.e., at least five fold increase), *Streptomycetaceae* and *Acidimicrobiaceae* (*Actinobacteria*), as well as taxa that do not include any known 4-NP isolates such as other unclassified *Actinobacteria* (Supplementary Figure S9). This finding agreed with those of previous studies (Zhang, Liu et al. 2009) (Liu, Zhang et al.

2005) that identified *Pseudomonas* sp. strain WBC-3 as 4-NP degrader, sharing ≥ 98% 16S rRNA gene sequence identity to our isolates. Notably, we have observed (uncharacterized) *Actinobacteria* that increased in their abundance during 4-NP perturbation, which echoed several previous studies (Jain, Dreisbach et al. 1994, Schäfer, Harms et al. 1996, Chauhan, Chakraborti et al. 2000). However, the representative 16S sequence of the *Actinobacteria* OTU showed only 83% nucleotide identity to the previously described 4-NP isolates (i.e., *Arthrobacter protophormiae*), indicating our 4-NP degraders belong to a novel *Actinobacteria* clade.

To date, only about 35 strains have been isolated and experimentally characterized as caffeine degraders that belong to phylogenetically diverse taxa of bacteria (Summers, Mohanty et al. 2015), such as *Rhodococcus* and *Pseudomonas* sp. CBB1. These strains degrade caffeine via C-8 oxidation pathways, and *Pseudomonas* sp. CES, *Pseudomonas* sp. CBB5 (Yu, Louie et al. 2009, Yu, Summers et al. 2014), and *Serratia marcescens* degrade caffeine via *N*-demethylation pathways (Mazzafera, Olsson et al. 1996). Several additional taxa such as *Moraxella* sp*., Alcaligenes* sp., (Mohapatra, Harris et al. 2006), and *Coryneform* have been identified as caffeine degraders previously (Madyastha and Sridhar 1998, Yamaoka-Yano and Mazzafera 1998) but their caffeine degradation pathways remain uncharacterized. For the caffeine mesocosms (Supplementary Figure S9), the abundance of *Comamonadaceae* (*Betaproteobacteria*) substantially increased (i.e., 4-fold more abundant), and its 16S rRNA gene sequence identity to our isolates was greater than 99%. Additional OTUs that increased in abundance in the post-enrichment metagenome included *Burkholderiaceae*

(*Betaproteobacteria*), uncultivated members of *Acidimicrobiaceae* (*Actinobacteria*), and *Proteobacteria* (showing <85% 16S rRNA gene identity to any cultured taxon in Greengene database). Therefore, it appears that several caffeine degraders, which escaped cultivation based on our efforts, may also have contributed to the entire microbial community. More notably, *Burkholderiaceae* has not been reported as caffeine degraders yet [reviewed in (Summers, Mohanty et al. 2015)], indicating that the corresponding OTUs might represent novel caffeine degraders (or be associated with bi-products).

To compare community gene content shifts in the 2,4-D, 4-NP and caffeine metagenomes before and after perturbation, we determined the relative coverage (X) of each protein-coding gene and compared the predicted gene functions based on Gene Ontology (GO) terms (Ashburner, Ball et al. 2000). For 2,4-D (Figure 8), the most enriched genes (i.e., showing > 6-fold difference in abundance relative to T=0) were associated with: 1) cell motility, 2) energy generation and maintenance, 3) transporters, 4) viral functions, and 5) several genes associated with 2,4-D biodegradation, e.g., catechol-1, 2 dioxygenase (19-fold difference), hydroxyquinol 1,2-dioxygenase (14-fold), 2,4-dichlorophenol 6-monooxygenase (11-fold) and chlorocatechol 1,2-dioxygenase (27-fold). The 2,4-D mesocosm II and III exhibited similar profiles except that enriched genes also included genes encoding proteins involved in phosphate metabolism, which indicated that phosphorus might have been limiting in these mesocosms. Overall, however, the majority of gene functions did not changed in abundance, e.g., 70.7% of total functions detected changed by less than two fold in abundance between T=0 and T=5.

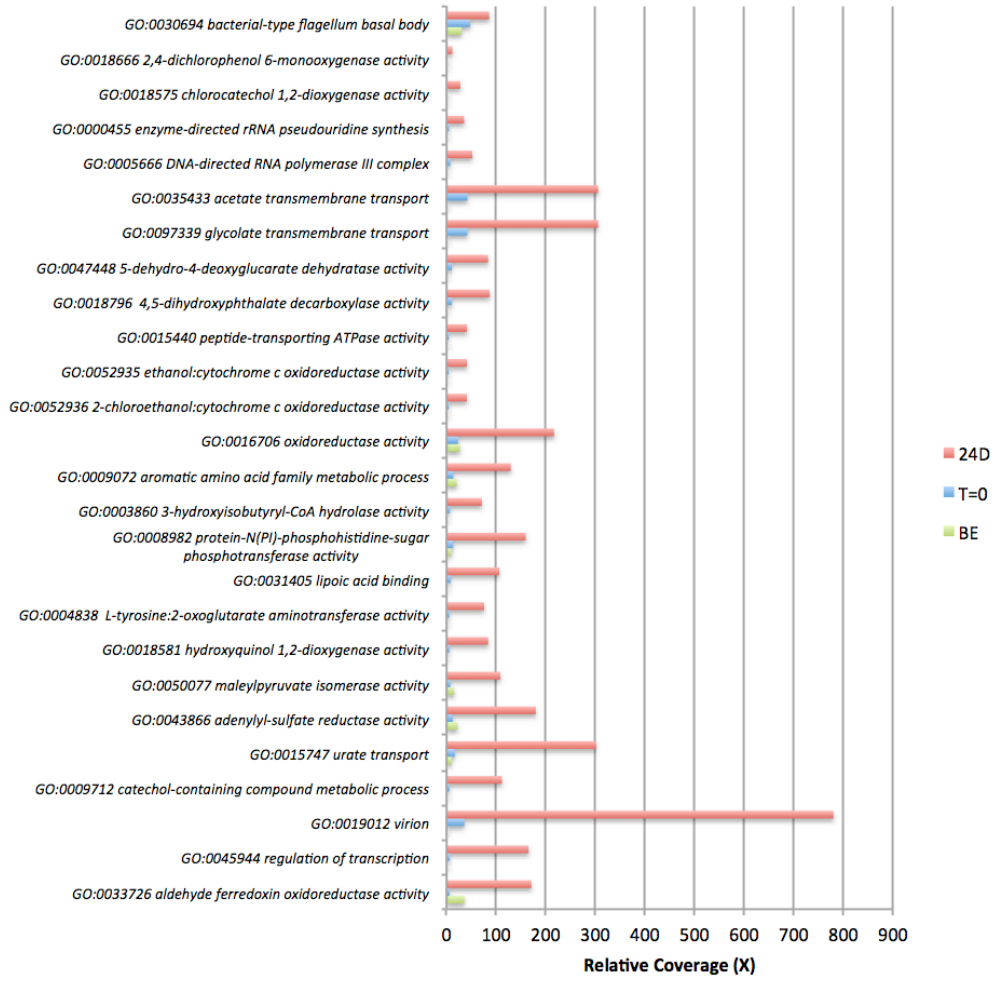

Figure 8. **Community gene content shifts in 2,4-D mesocosm I as an effect of 2,4-D addition**. The relative coverage (X; x-axis) of genes (y-axis) was calculated by summing the length of all reads mapping on the gene (a minimum cut-off for a match of ≥80 bp alignment length and ≥97% nucleotide identity) and dividing it by the length of the gene sequence. The most differentially abundant genes between T=0 and T=5 metagenomes were clustered by their GO terms and are shown on the graph. The GO accession number is also provided, followed by the GO description. BE: Bottle effect sample (control).

For 4-NP and caffeine (Supplementary Figure S10), several of the enriched genes were similar to those for 2,4-D such as bacterial-type flagellum, transmembrane transport activity and viral activities. Enriched genes also included 1) phenol catabolic process (6-fold), reflecting 4-NP biodegradation bi-products; 2) spore germination (54-fold),

probably due challenging growth conditions; and 3) caffeine-related catabolic pathways, including urate/purine-containing compounds metabolism (39-fold), xanthine oxidase/dehydrogenase (6-fold) and demethylase activity (1.7-fold more abundant). Further, genes associated with maintenance of CRISPR repeat elements and membrane signal transduction activity showed 5-fold and 8-fold higher abundance in caffeine and 4-NP mesocosms, respectively, reflecting likely active viral predation.

The results reported above were presumably attributable to the selection pressure applied by the addition of the organic compound and were not related to bottle effects from the incubation. Consistent with this interpretation, we observed that the datasets from the bottle effect control (BE), i.e., lake water incubated without the addition of the substrate, resembled the T=0 (initial) community much more than those from the final sampling point for the mesocosms amended with the three compounds used in this study, both in terms of taxa composition and gene functions. For instance, the BE datasets had a small decrease, if any, in the number of OTUs detected relative to the T=0 dataset (compared to about 35% of the OTUs becoming undetectable in the 2,4-D mesocosms) (Table S3), and no bacteriophage genes were strongly enriched in BE vs. T=0 datasets (Figure 8, S10, S11).

CHAPTER 4

DISCUSSION

The three 2,4-D mesocosms showed substantially different biodegradation profiles of the added compounds. Sequencing of time-series samples revealed that these profiles were presumably due to variation in the pool of rare taxa in the inoculum for each mesocosm, even though the inocula originated from the same mixed (homogenized) lake water collected in a single sampling trip to Lake Lanier. All organisms enriched during the incubation period represented rare initial populations *in-situ*, e.g., below the detection limit of metagenomic efforts in the T=0 sample. These populations were usually assignable to different species in each replicate mesocosm, and encoded distinct versions of the biodegradation genes. For instance, we found that the microbial communities of the 2,4-D mesocosms II and III, which showed slow degradation, encoded different degradation genes than those in mesocosm I (exhibiting robust degradation), including several homologs of *tfdAα*, which typically show weak 2,4-D degradation (Itoh, Kanda et al. 2002). Addition of nutrients (e.g., ¼ – strength MSB liquid media, a source of nitrogen as well as other nutrients) to small volumes (microcosms) originating from mesocosms II and III greatly increased the degradation rate exhibited by these populations. These findings show that not only were different taxa enriched in each mesocosm but also it was likely that each of the taxa found in the different mesocoms have different nutrient requirements and encode genes that have different biodegradation kinetics and/or regulation. Collectively, these findings showed that the rare biosphere enabled microbial community adaption to the changing conditions

(i.e., promote the degradation of the added organic compounds); however, the specific taxa and genes involved (i.e., enriched during this process) were highly variable. Thus, stochastic processes in the distribution and gene regulation of these taxa apparently played a key role during community adaptation, while the pool of rare species and taxa contributing to adaptation is large, at least based on the 2,4-D mesocosms. These findings were also consistent with recent studies that have highlighted the importance of stochastic processes, e.g., variation in abundance, extinction, speciation and evolution, in mediating microbial community response to perturbations [e.g., (Prach and Walker 2011, Zhou, Deng et al. 2014)].

As the mesocosms were perturbed with 2,4-D, the microbial community was under strong selection pressure. The organisms that were capable of using the 2,4-D as a sole carbon and energy source for growth became more abundant and gained an advantage compare to other populations. For 2,4-D degraders, their degradation genes were often found on a large plasmid (e.g., *Burkholderia* sp. KK1 plasmid, ~580kb). Annotation of the KK1 plasmid (Supplementary Table S4) showed that it encodes all necessary *tra* genes for transmission except for *traI* gene that encodes the relaxase protein. Comparison of the KK1 plasmid against those available in the public databases showed that it represents a fusion of plasmid pM7012 (Sakai, Ogawa et al. 2014) and pJp4 (Don and Pemberton 1981, Trefault, De la Iglesia et al. 2004) reported previously (Supplementary Figure S4), showing very high nucleotide identities to the latter plasmids for the regions in common (>99% nucleotide identity). The latter plasmids have been shown to be mobile (Sakai, Ogawa et al. 2014). In addition, the plasmid *tfd* gene clusters

were flanked by repeating sequences and sequences encoded phage integrases and transposases, which suggests that the plasmid and its *tfd* gene clusters have undergone intensive genetic recombination events recently. Finally, the plasmid to chromosome ratio in strain *Burkholderia* sp. KK1, grown in isolation and harvested and sequenced at the mid-to-late exponential phase, was ~1.4, while the same ratio in the mesocosm increased over time: T=0 (N/A), T=2 (1.30), T=4 (2.0) and T=5 (2.9). These findings indicate that other members of the community might have acquired the plasmids during the incubation. The alternative hypothesis that the plasmid copy number increased due to selection pressure does not appear to be as parsimonious because the ratio after the first 2,4-D spike in event was close to the pure isolate ration (1.3 vs. 1.4). Therefore, it is intriguing to hypothesize that the KK1 plasmid is also mobile, and it was transferred horizontally during the incubation period, between co-occurring members of the community, due to strong selection pressure and thus, mediated microbial community adaptation to the added 2,4-D. However, whether or not such a molecular mechanism played an important role during the incubation awaits experimental verification.

In addition to degradation genes, the relative abundance of several genes associated with phage proteins, flagella motility, chemotaxis, and conjugative transfer increased in post-enrichment samples relative to time 0 or the bottle effect datasets (Figure S11). These results are likely attributable to the stressful environmental conditions, e.g., toxicity of added organic compounds, lack of nutrients, and bacteriophage predation. Consistent with these interpretations, about ~35% of all OTUs decreased in abundance at the end of each mesocosm incubation but not in the dataset

from the bottle effect incubation (Table S3). Under such conditions, lysogenic bacteriophages have been shown to transition to a lytic phase (Middelboe, Jorgensen et al. 1996, Wells and Deming 2006, Ram and Sime-Ngando 2008, Payet and Suttle 2013). Moreover, the patterns observed may also be attributed, at least partly, to "kill the winner" scenario (Thingstad and Lignell 1997). Under this scenario, the most successful population, e.g., the one with efficient biodegradation genes under the growth conditions of our mesocosms is preferentially targeted by bacteriophages. The decrease of the abundant *Burkholderia* sp. KK1-like population at the last sampling point after exponential growth in the three preceding sampling points that followed the addition of 2,4-D (Figure 4) is consistent with the patterns expected under the "kill the winner" scenario. Cell motility is also expected to increase due to the nutrient limiting conditions and the lack of continuous stirring of our mesocosms.

Another interesting observation was that at least three distinct *tfd* operons (Figure 5) and several abundant OTUs/population co-existed after the addition of 2,4-D, without any one population outcompeting the remaining ones apparently. These results could be attributable to differential bacteriophage predation of the abundant populations or to different kinetics of the Tfd enzymes, e.g., enzymes with high vs. low affinity for the substrate. The former scenario has been well documented several times previously (Fuhrman and Schwalbach 2003, Escobar‐Paramo, Faivre et al. 2009, Rodriguez-Valera, Martin-Cuadrado et al. 2009, Weitz and Wilhelm 2012). However, whether or not this scenario applies to our mesocosm incubations remains to be experimentally determined.

Overall, our results strongly support the hypothesis that the rare organisms and genes serve as a genetic reservoir that can contribute to whole microbial community response to environmental perturbations such as the addition of organic pollutants or toxins.

APPENDIX A

SUPPLEMENTARY FIGURES

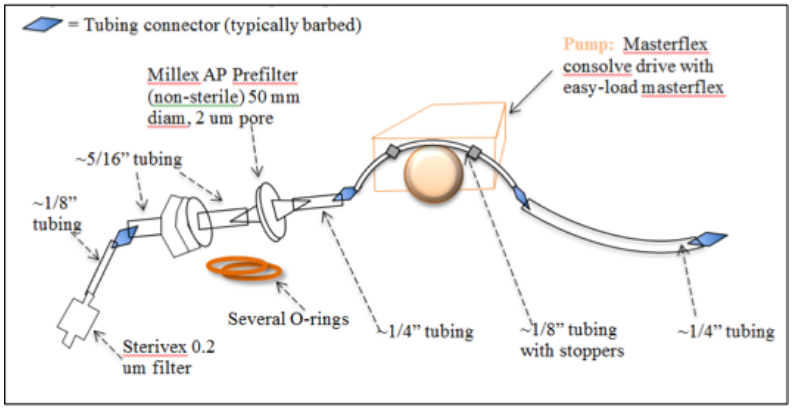

Figure S1. **overview of filtration system.** First, the AP filters were used, followed by GF/A (1.6 mm, Whatman, Pittsburgh, USA) and Sterivex filters (0.22 mm, Millipore, Fisher Scientific, Canada).

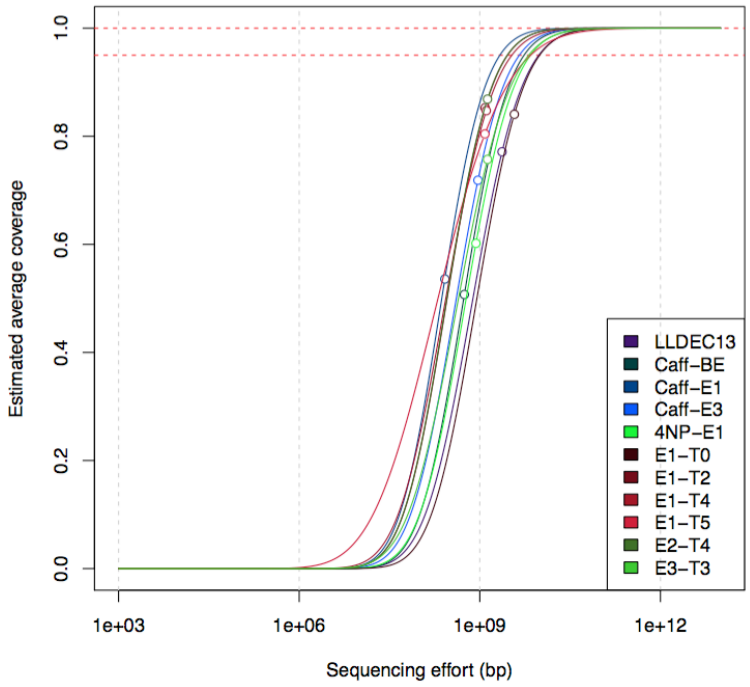

Figure S2. **Complexity of 2,4-D, 4-NP and caffeine mesocosm microbial communities as assessed by Nonpareil curves.**

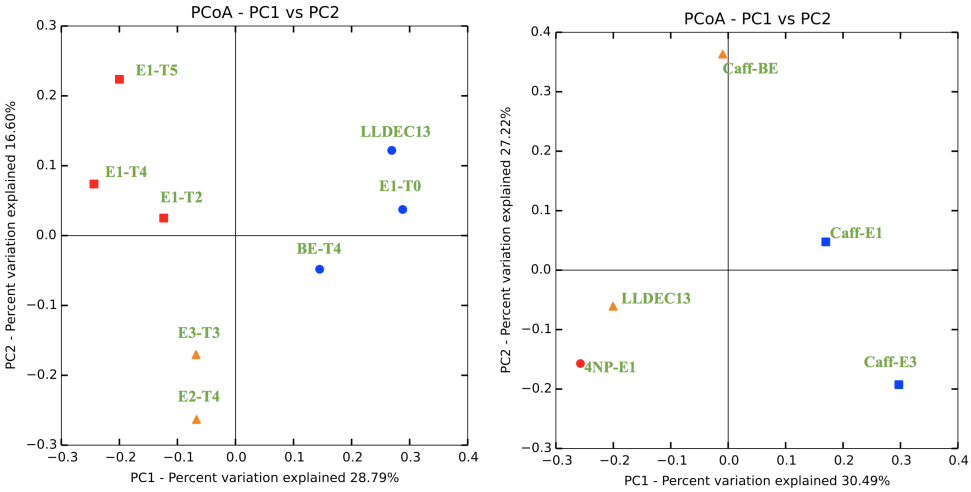

Figure S3. **Comparison of microbial community composition among 2,4-D (upper part), 4-NP and caffeine (lower part) mesocosms.** Beta-diversity similarities among the communities were calculated using binary_sorensen_dice metrics as implemented in QIIME**.**

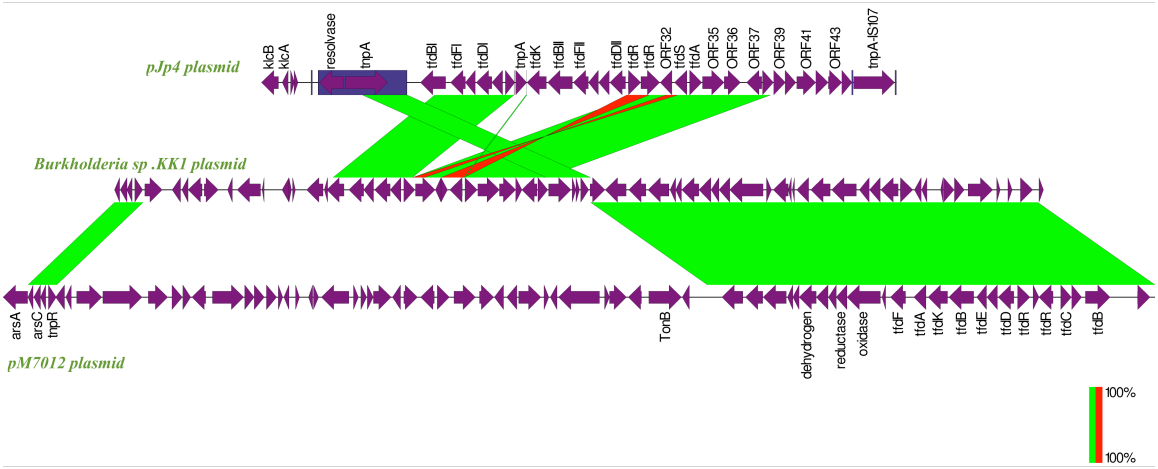

Figure S4. **Comparison of the *tfd* genes clusters among KK1 plasmid, pJp4 and pM7012 plasmids.** The *tfd* genes and flanking genes are shown in purple color and mobile elements in blue background. Lines connecting the plasmid represent 100% nucleotide identity; green denoting same orientation and red denoting opposite orientation.

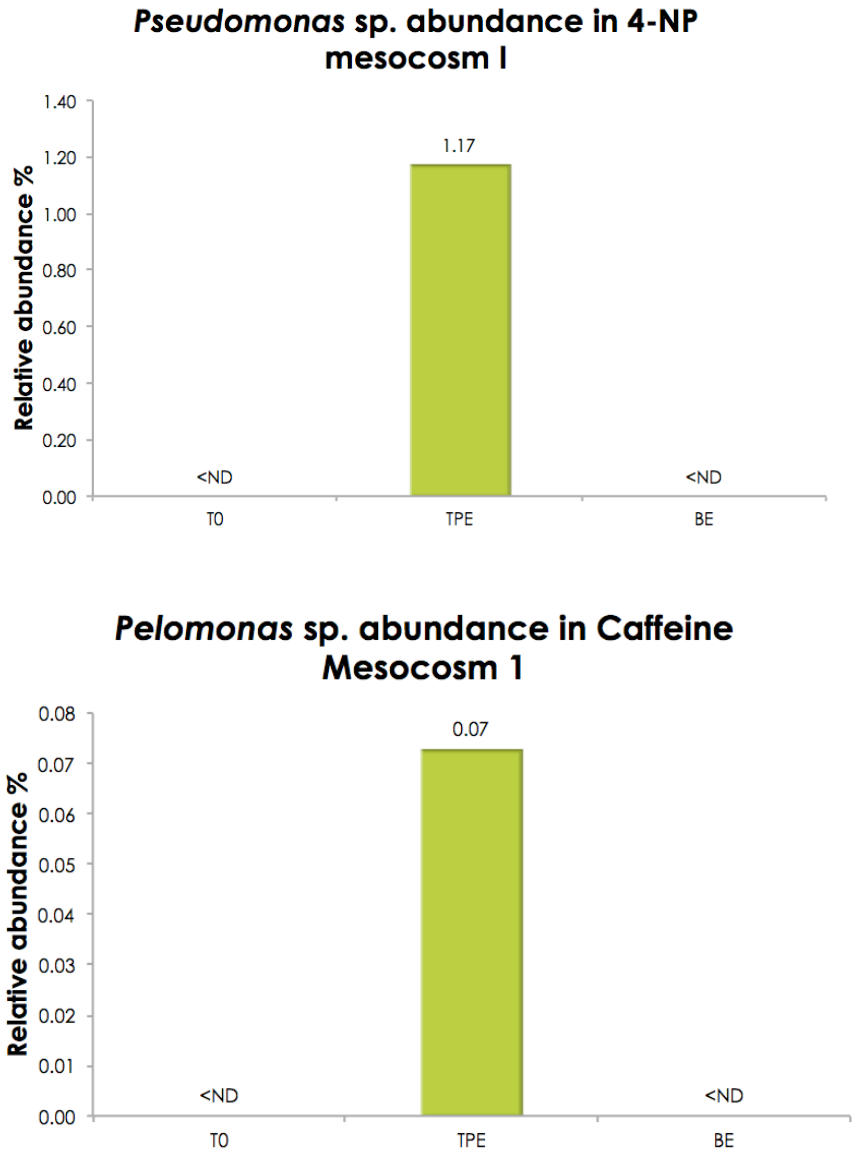

Figure S5. **Relative abundance of 4-NP and caffeine degraders in the mesocosm they originated form.** Note that both isolates were rare at T = 0, but became abundant at the last time point (TPE), similar to the 2,4-D degrader profiles shown in Figure 3. BE: Bottle effect sample (control).

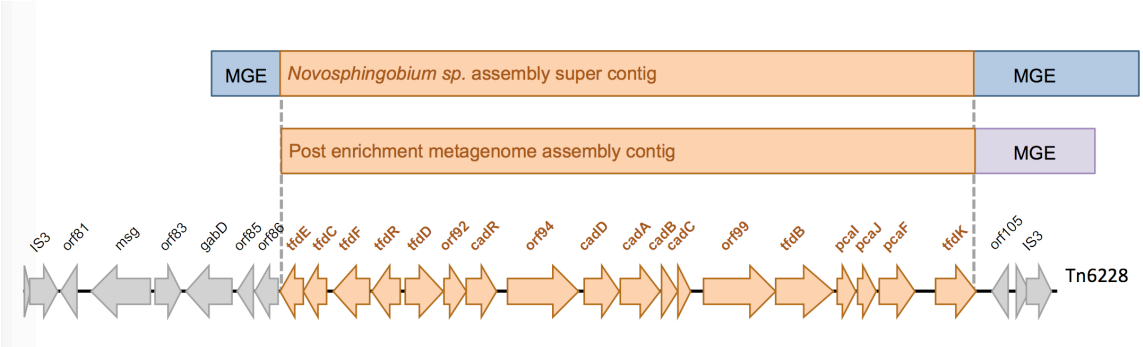

Figure S6. **Alignment of *cad* genes from *Novosphingobium* sp. 24J isolate and metagenomic contig against that of *Sphingobium* sp. ERG5 Transposon Tn6288.** *Sphingobium* sp. ERG5 Transposon Tn6288 (accession number NC_022235) (Nielsen, Xu et al. 2013)

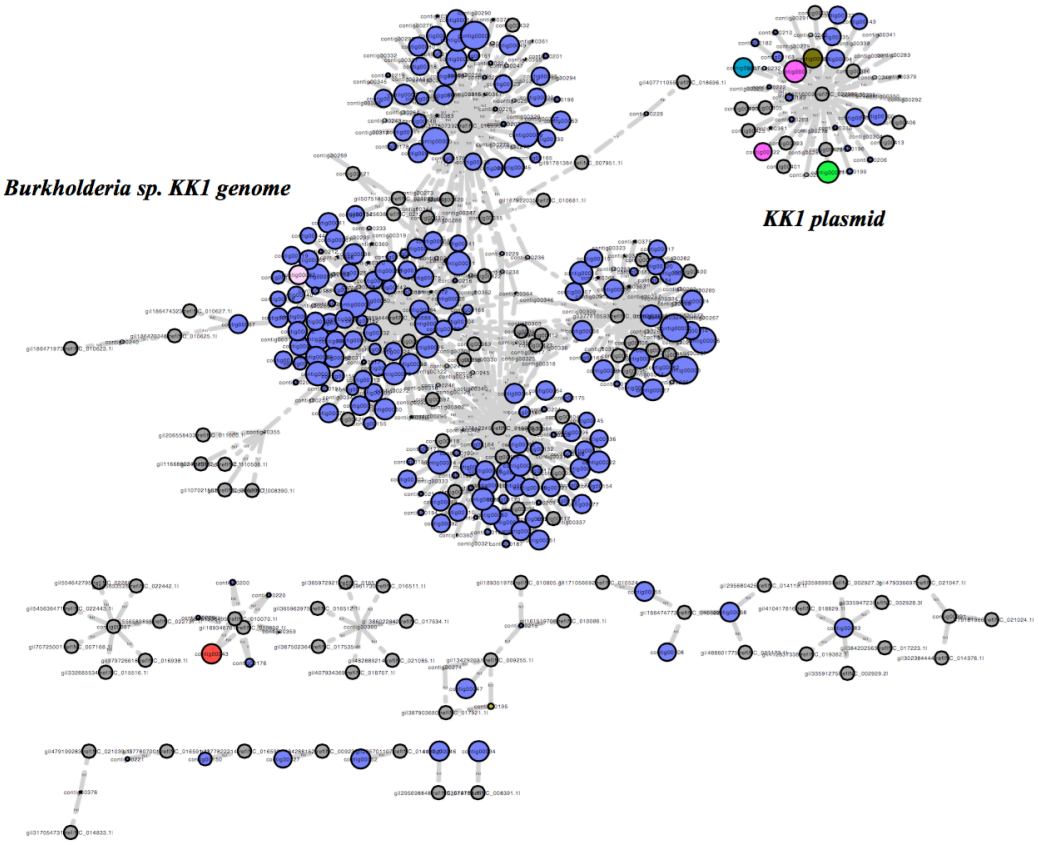

Figure S7. **Verification of scaffolding *Burkholderia* sp. KK1 genome and KK1 plasmid using PLACNET.** The KK1 plasmid is shown on upper right. Red: contigs with relaxase (conjugation required); Yellow: contigs encoded protein involved in T4SS system; Dark Brown: *arsA* (arsenic resistance) and 2,4-D genes; Green: contigs encoded *parAB*, *repA* and RIP (plasmid initiator replication); Purple: tRNA regions; Cyan: *recD* (helicase). Blue: the size of contigs ≥ 200bp. The results confirmed that the identified scaffold represented a plasmid and indicated maybe two or more small plasmids co-exist in the KK1 genome.

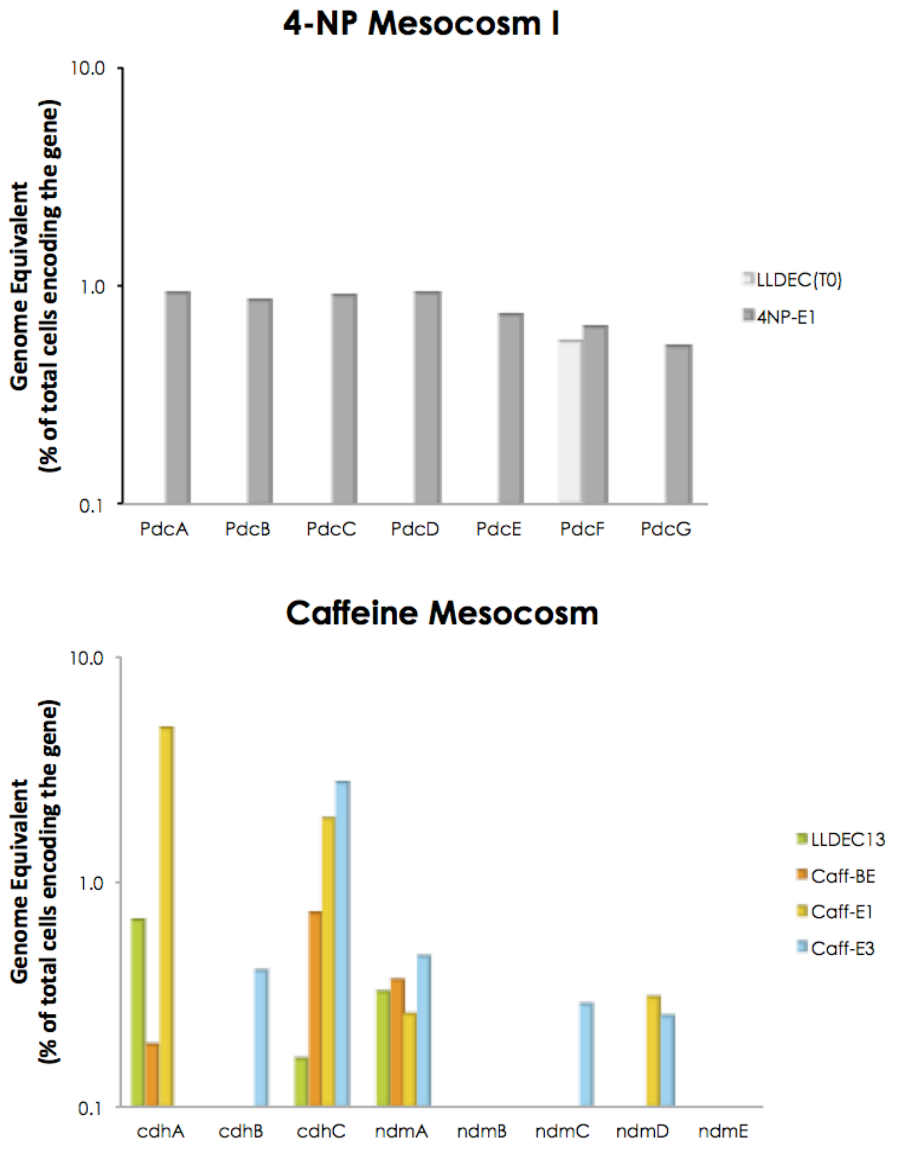

Figure S8. **Abundance of 4-NP degradation genes *pdcABCDEFG* and caffeine degradation genes *cdhABC* and *ndmABCDE* in our mesocosms**. Abundance was estimated as average genome equivalents (% of total cells encoding the gene) at each time point in corresponding mesocosms. Note: for caffeine degradation, the identified genes in the mesocosm metagenomes showed only moderate relatedness to previously described caffeine degradation genes, e.g., ~50% amino acid identity with *cdhABC* genes (Summers et al. 2013) and *ndmACD* genes (Mohanty et al. 2012).

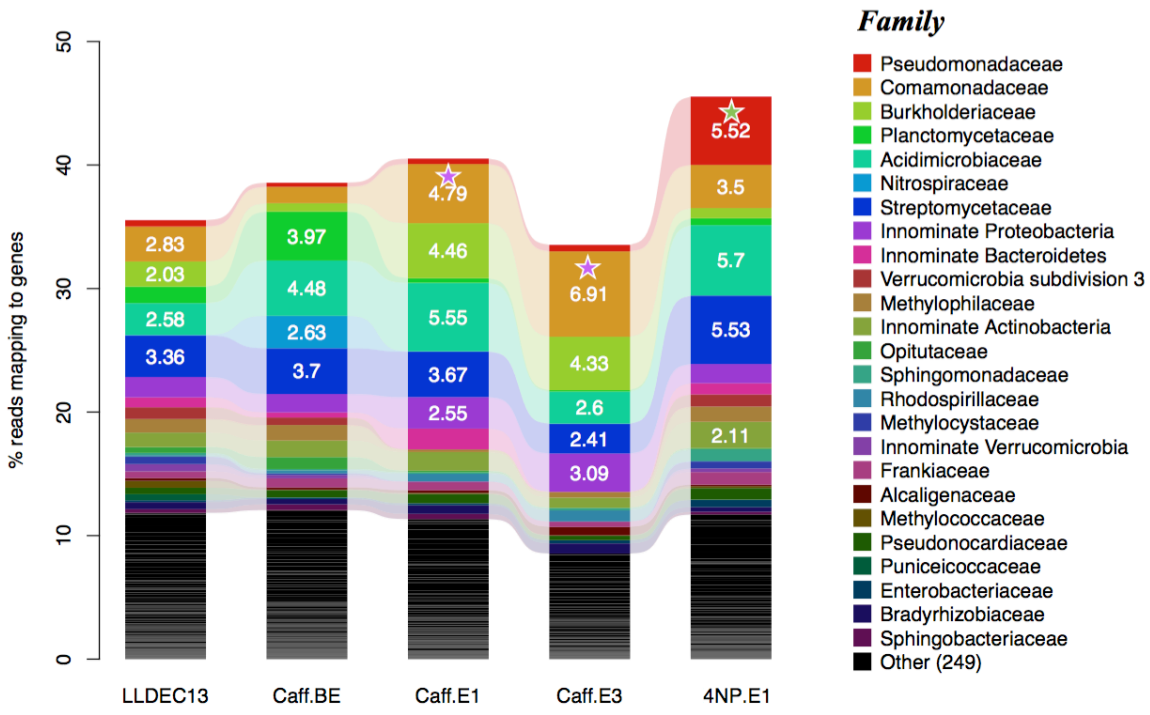

Figure S9. **Taxonomy shifts in 4-NP mesocosm I at T = 2, caffeine mesocosm I, III at T=2 and bottle effect (BE), and the original lake (LLDEC13).** The purple asterisk denotes the family that *Pelomonas* sp. belongs to in Caff-E1 and E3 based on 16S rRNA gene identity (100% nucleotide identity cutoff). The green asterisk denotes the family that *Pseudomonas* sp. belong to in 4NP-E1 (100% nucleotide identity cutoff).

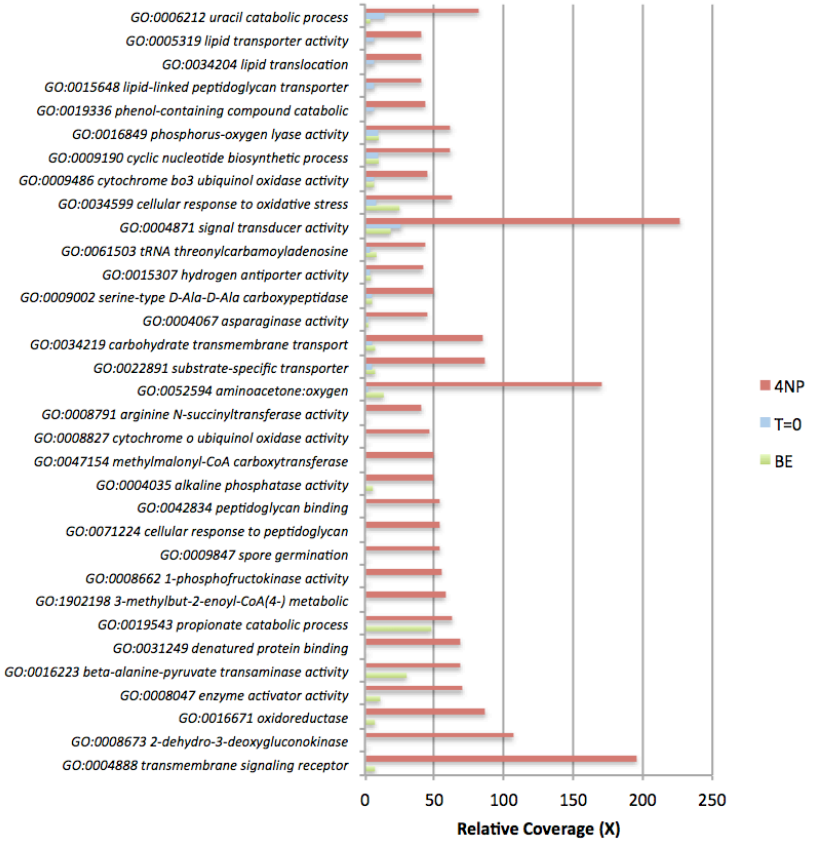

Figure S10. **Community gene content shifts in 4NP (upper part) and caffeine (lower part) mesocosms as an effect of 4-NP and caffeine addition, respectively**. The relative coverage (X; x-axis) of genes (Y; y-axis) was calculated by summing the length of all reads mapping on the gene (a minimum cut-off for a match of ≥80 bp alignment length and ≥97% nucleotide identity) and dividing it by the length of the gene sequence. The most differentially abundant genes between T=0 and T=2 metagenomes were clustered by their GO terms and are shown on the graph. The GO accession number is also provided, followed by the GO description. BE: Bottle effect sample (control).

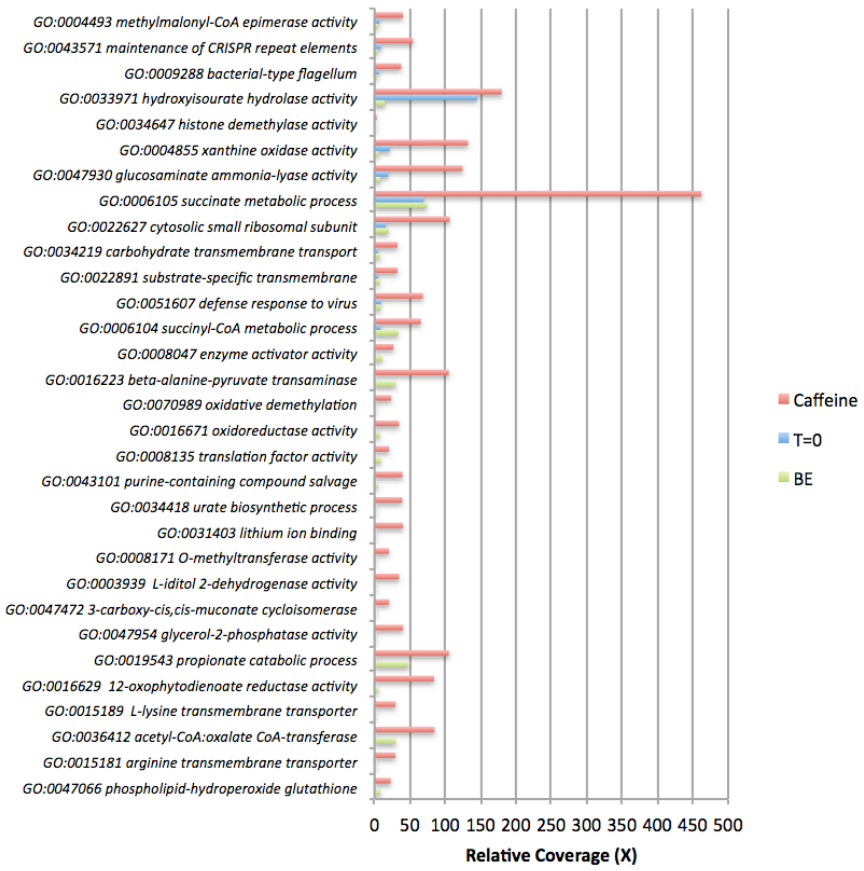

Figure S10 continued.

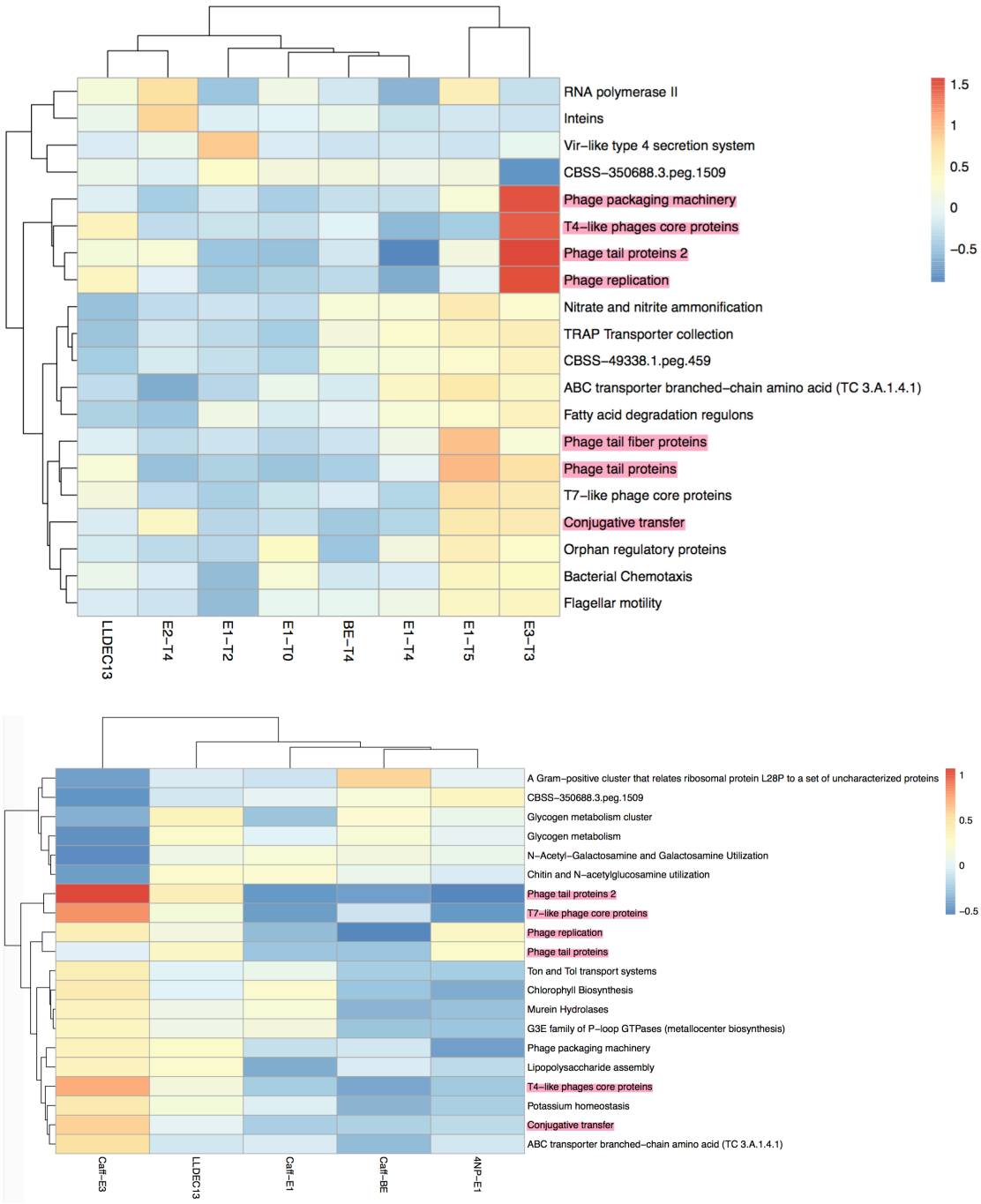

Figure S11. **Functional gene content shifts in the 2,4-D (upper part), 4-NP, and caffeine (lower part) mesocosms.** The protein-coding genes were annotated using the SEED subsystems database and clustered based on changes in their abundance relative to T=0 (i.e., 2,4-D, 4NP and caffeine compared to reference level LLDEC13). Clustering was performed using SEED subsystem I and the most differentially abundant genes based on parallel analysis as implemented in the DESeq2 package (Love, Huber et al. 2014) are shown.

APPENDIX B

SUPPLEMENTARY TABLES

Table S1 **2,4-D Mesocosms DNA sampling strategy**

Mesocosm I

Sample Name	Total Volume	*Volume per Sterivex*	Sterivex No.	The filters used	Date of Sampling	Number of Sterivex
Mesocosm T0	5L	2.5L	T0-1 T0-2	GF/A, sterivex	12/26/13	2
EXP1 T1	1L			GF/A, sterivex	12/30/13	1
BE T1	1L			GF/A, sterivex	12/30/13	1
EXP1 T2	5L	2.5L	T2-1 T2-2	GF/A, sterivex	1/4/14	2
BE T2	5L	2.5L	BET2-1 BET2-2	GF/A, sterivex	1/4/14	2
EXP1 T3	1L			GF/A, sterivex	1/6/14	1
BE T3	1L			GF/A, sterivex	1/6/14	1
EXP1 T4	5L	2.5L	T4-1 T4-2	GF/A, sterivex	1/7/14	2
BE T4	5L	2.5L	BET4-1 BET4-2	GF/A, sterivex	1/7/14	2
EXP1 T5	4.6L	4.6L	T5	GF/A, sterivex	1/12/14	1
BE T5	5L	5L	BET5	GF/A, sterivex	1/12/14	1
					Sum	16

EXP1 stands for 2,4-D Mesocosm EXP1;

BE stands for Bottle Effect. Exp1 and BE sampled simultaneously.

T=0 DNA sample is all the same for three mesocosms.

Highlighted samples have been sequenced by Illumina HiSeq.

Mesocosm II

Sample Name	Total Volume	Sterivex No.	The filters used	Date of Sampling	Number of Sterivex
T0	5L		GF/A, sterivex	12/26/13	0
EXP2 T1	1L		GF/A, sterivex	12/30/13	1
EXP2 T2	1L		GF/A, sterivex	1/12/14	1
EXP2 T3	1L		GF/A, sterivex	1/30/14	1
EXP2 T4	1L	2-T4	GF/A, sterivex	2/8/14	1
				Sum	4

Mesocosm III

Sample Name	Total Volume	Volume per Sterivex	Sterivex No.	The filters used	Date of Sampling	Number of Sterivex
T0	5L			GF/A, sterivex	12/26/13	0
EXP3 T1	1L			GF/A, sterivex	12/30/13	1
EXP3 T2	1L			GF/A, sterivex	1/12/14	1
EXP3 T3	5L		T3-1, T3-2, T3-3	GF/A, sterivex	1/16/14	3
					Sum	5

Table S2 **Trimmed length and assembly statistics for each metagenome**

	Trimmed reads length	Total number of trimmed reads	Average contigs size	N50 contig size	Largest Contig size	Number of large Contig (>500bp)
LLDEC13	131.203	18220737	1029	1062	54874	173541
2,4-D mesocosm I at T=0 (E1-T0)	131.16	29145453	1158	1254	58554	80221
2,4-D mesocosm I at T=2 (E1-T2)	133.787	9361741	1158	1254	58554	80221
2,4-D mesocosm I at T=4 (E1-T4)	136.583	9600373	1067	1123	47834	92891
2,4-D mesocosm I at T=5 (E1-T5)	131.699	9417412	1292	1506	68557	73753
2,4-D mesocosm II at T=4 (E2-T4)	134.059	9495080	1250	1448	40480	83309
2,4-D mesocosm III at T=3 (E3-T3)	134.86	10212412	1129	1224	47767	92898
BE mesocosm at T=4 (BE-T4)	135.999	5934275	960	979	20554	55695
Caffeine Mesocosm I at T=2 (Caff-E1)	212.33	1252602	941	932	44283	54248
4-NP mesocosm I at T=2 (4NP-E1)	189.511	4481582	880	872	46465	136637

Table S3 **alpha-diversity of 2,4-D, 4-NP and caffeine mesocosms**

2,4-D mesocosms

	observed_otus	shannon	PD_whole_tree
LLDEC	515	8.17	47.92
BE-T4	442	7.89	41.03
E1-T0	511	8.08	46.81
E1-T2	340	7.18	33.66
E1-T4	313	6.84	31.83
E1-T5	339	6.93	29.60
E3-T3	430	7.66	36.91
E2-T4	445	7.91	39.93

4-NP and caffeine mesocosms

	observed_otus	shannon	PD_whole_tree
LLDEC	515	8.17	47.92
Caff-BE	491	8.14	50.78
Caff-E1	377	7.52	37.56
Caff-E3	350	7.25	31.87
4NP-E1	450	7.77	44.27

Table S4 ***Burkholderia*** **sp. KK1 plasmid annotation**

contig00035	gene_479|GeneMark.hmm|1230_nt|-|356073|357302	TfdB
contig00035	gene_480|GeneMark.hmm|1065_nt|-|357647|358711	maleylacetate reductase(TfdF)
contig00035	gene_481|GeneMark.hmm|705_nt|-|358708|359412	dienelactone hydrolase(TfdE)
contig00035	gene_482|GeneMark.hmm|1113_nt|-|359487|360599	chloromuconate cycloisomerase(TfdD)
contig00035	gene_483|GeneMark.hmm|768_nt|-|360596|361363	chlorocatechol -dioxygenase(TfdC)
contig00035	gene_484|GeneMark.hmm|795_nt|plus|361574|362368	TfdT
contig00455	gene_485|GeneMark.hmm|1392_nt|plus|362414|363805	regulatory protein
contig00455	gene_486|GeneMark.hmm|402_nt|-|363833|364234	family transcriptional regulator
contig00474	gene_487|GeneMark.hmm|450_nt|plus|364236|364685	regulatory protein
contig00209	gene_488|GeneMark.hmm|888_nt|-|364858|365745	TfdR
contig00209	gene_489|GeneMark.hmm|864_nt|plus|365951|366814	TfdA_pJp4
contig00209	gene_490|GeneMark.hmm|1542_nt|plus|366857|368398	acyl- synthetase
contig00209	gene_491|GeneMark.hmm|1167_nt|plus|368404|369570	thiolase
contig00209	gene_492|GeneMark.hmm|441_nt|plus|369567|370007	Protein of unknown function DUF35
contig00209	gene_493|GeneMark.hmm|1077_nt|-|370004|371080	vanillate o-demethylase oxygenase chain a
contig00209	gene_494|GeneMark.hmm|705_nt|plus|371155|371859	transcription regulator
contig00209	gene_495|GeneMark.hmm|1593_nt|plus|371916|373508	transposase tn3 family protein
contig00209	gene_496|GeneMark.hmm|270_nt|plus|373598|373867	hypothetical protein BCh11DRAFT_05164
contig00209	gene_497|GeneMark.hmm|264_nt|plus|373864|374127	hypothetical protein Reut_D6458
contig00209	gene_498|GeneMark.hmm|573_nt|plus|374204|374776	MULTISPECIES: hypothetical protein
contig00209	gene_499|GeneMark.hmm|1059_nt|plus|374873|375931	mfs transporter
contig00209	gene_500|GeneMark.hmm|1494_nt|-|375953|377446	dsba oxidoreductase
contig00209	gene_501|GeneMark.hmm|1224_nt|-|377640|378863	dienelactone hydrolase
contig00200	gene_502|GeneMark.hmm|1491_nt|-|379016|380506	amp-dependent synthetase and ligase
contig00200	gene_503|GeneMark.hmm|396_nt|-|380588|380983	endoribonuclease l-psp
contig00200	gene_504|GeneMark.hmm|420_nt|-|380980|381399	4-hydroxybenzoyl thioesterase
contig00200	gene_505|GeneMark.hmm|1224_nt|-|381420|382643	acyl- dehydrogenase
contig00200	gene_506|GeneMark.hmm|852_nt|-|382645|383496	enoyl- hydratase
contig00200	gene_507|GeneMark.hmm|561_nt|-|383516|384076	family transcriptional regulator
contig00200	gene_508|GeneMark.hmm|774_nt|-|384073|384846	short-chain dehydrogenase reductase sdr
contig00200	gene_509|GeneMark.hmm|2385_nt|-|384843|387227	nadh: flavin oxidoreductase nadh oxidase
contig00200	gene_510|GeneMark.hmm|354_nt|plus|387465|387818	hypothetical protein
contig00200	gene_511|GeneMark.hmm|1071_nt|-|387948|389018	maleylacetate reductase(TfdF)
contig00200	gene_512|GeneMark.hmm|222_nt|-|389018|389239	hypothetical protein
contig00200	gene_513|GeneMark.hmm|159_nt|-|389310|389468	2-keto-4-pentenoate hydratase
contig00200	gene_514|GeneMark.hmm|864_nt|-|389616|390479	TfdA_pM7012
contig00200	gene_515|GeneMark.hmm|1392_nt|-|390631|392022	TfdK
contig00200	gene_516|GeneMark.hmm|1764_nt|-|392121|393884	TfdB
contig00445	gene_517|GeneMark.hmm|708_nt|-|394098|394805	dienelactone hydrolase(TfdE)
contig00445	gene_518|GeneMark.hmm|762_nt|-|394826|395587	chlorocatechol -dioxygenase(TfdC)
contig00355	gene_519|GeneMark.hmm|1236_nt|-|395637|396872	chloromuconate cycloisomerase(TfdD)
contig00355	gene_520|GeneMark.hmm|888_nt|plus|397025|397912	TfdR
contig00355	gene_521|GeneMark.hmm|483_nt|-|398037|398519	transposase IS66
contig00355	gene_522|GeneMark.hmm|378_nt|-|398547|398924	TfdS
contig00355	gene_523|GeneMark.hmm|171_nt|-|399905|400075	chloromuconate cycloisomerase(TfdD)
contig00445	gene_524|GeneMark.hmm|762_nt|plus|400108|400869	chlorocatechol -dioxygenase(TfdC)
contig00445	gene_525|GeneMark.hmm|708_nt|plus|400890|401597	dienelactone hydrolase(TfdE)
contig00388	gene_526|GeneMark.hmm|1761_nt|plus|401867|403627	TfdB

Table S5 **qPCR primers**

	Primer sequences
Forward primer tfdA-F	5’- CTC GAA GGC GGT TTC ATC A -3’
Reverse primer tfdA-R	5’ -GTT GAT TCG CGA AGT TCC C -3’
Forward primer 1055yF	5’ -ATG GYT GTC GTC AGC T -3’
Reverse primer 1392R	5’ -ACG GGC GGT GTG TAC -3’

Table S6 **The statistics of each assembled isolates genome**

Genome	No. of basepairs	No. of contigs	Average contig size	N50 contig size	Largest contig size	Completeness	Contamination	Strain heterogeneity
Burkholderia	9,517,307	583	19,304	36,095	170,704	100	0	0
Sphingopyxis	4,333,434	309	14,024	24,828	85,219	98.28	0	0
Variovorax	7,177,164	197	36,432	68,108	217,155	100	0	0
Pseudomonas	6,374,739	262	24,331	49,190	164,660	98.28	0	0
Pelomonas	6,854,870	558	12,284	24,572	82,945	98.28	0	0

REFERENCES

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." Journal of molecular biology **215**(3): 403-410.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight and J. T. Eppig (2000). "Gene Ontology: tool for the unification of biology." Nature genetics **25**(1): 25-29.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell and D. L. Wheeler (2004). "GenBank: update." Nucleic acids research **32**(Database issue): D23.

Campbell, B. J., L. Yu, J. F. Heidelberg and D. L. Kirchman (2011). "Activity of abundant and rare bacteria in a coastal ocean." Proc Natl Acad Sci U S A **108**(31): 12776-12781.

Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich and J. I. Gordon (2010). "QIIME allows analysis of high-throughput community sequencing data." Nature methods **7**(5): 335-336.

Chauhan, A., A. K. Chakraborti and R. K. Jain (2000). "Plasmid-encoded degradation of p-nitrophenol and 4-nitrocatechol by *Arthrobacter protophormiae*." Biochemical and biophysical research communications **270**(3): 733-740.

Conesa, A., S. Götz, J. M. García-Gómez, J. Terol, M. Talón and M. Robles (2005). "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research." Bioinformatics **21**(18): 3674-3676.

Cox, M. P., D. A. Peterson and P. J. Biggs (2010). "SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data." BMC bioinformatics **11**(1): 485.

Curtis, T. P., W. T. Sloan and J. W. Scannell (2002). "Estimating prokaryotic diversity and its limits." Proceedings of the National Academy of Sciences **99**(16): 10494-10499.

Dash, S. S. and S. N. Gummadi (2006). "Catabolic pathways and biotechnological applications of microbial caffeine degradation." Biotechnology letters **28**(24): 1993-2002.

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu and G. L. Andersen (2006). "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB." Applied and environmental microbiology **72**(7): 5069-5072.

Don, R. and J. Pemberton (1981). "Properties of six pesticide degradation plasmids isolated from *Alcaligenes paradoxus* and *Alcaligenes eutrophus*." Journal of Bacteriology **145**(2): 681-686.

Dupont, C. L., D. B. Rusch, S. Yooseph, M.-J. Lombardo, R. A. Richter, R. Valas, M. Novotny, J. Yee-Greenbaum, J. D. Selengut and D. H. Haft (2012). "Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage." The ISME journal **6**(6): 1186-1199.

Escobar‐Paramo, P., N. Faivre, A. Buckling, C. GOUGAT‐BARBERA and M. Hochberg (2009). "Persistence of costly novel genes in the absence of positive selection." Journal of evolutionary biology **22**(3): 536-543.

Falkowski, P. G., T. Fenchel and E. F. Delong (2008). "The microbial engines that drive Earth's biogeochemical cycles." Science **320**(5879): 1034-1039.

Fuhrman, J. and M. Schwalbach (2003). "Viral influence on aquatic bacterial communities." The Biological Bulletin **204**(2): 192-195.

Goris, J., K. T. Konstantinidis, J. A. Klappenbach, T. Coenye, P. Vandamme and J. M. Tiedje (2007). "DNA-DNA hybridization values and their relationship to whole-genome sequence similarities." Int J Syst Evol Microbiol **57**(Pt 1): 81-91.

Haft, D. H., J. D. Selengut, L. M. Brinkac, N. Zafar and O. White (2005). "Genome Properties: a system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics." Bioinformatics **21**(3): 293-306.

Hoffmann, D., S. Kleinsteuber, R. H. Müller and W. Babel (2003). "A transposon encoding the complete 2, 4-dichlorophenoxyacetic acid degradation pathway in

the alkalitolerant strain *Delftia acidovorans* P4a." Microbiology **149**(9): 2545-2556.

Hogan, D., D. Buckley, C. Nakatsu, T. Schmidt and R. Hausinger (1997). "Distribution of the tfdA gene in soil bacteria that do not degrade 2, 4-dichlorophenoxyacetic acid (2, 4-D)." Microbial ecology **34**(2): 90-96.

Huse, S. M., D. M. Welch, H. G. Morrison and M. L. Sogin (2010). "Ironing out the wrinkles in the rare biosphere through improved OTU clustering." Environmental microbiology **12**(7): 1889-1898.

Hyatt, D., G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer and L. J. Hauser (2010). "Prodigal: prokaryotic gene recognition and translation initiation site identification." BMC bioinformatics **11**(1): 119.

Itoh, K., R. Kanda, Y. Sumita, H. Kim, Y. Kamagata, K. Suyama, H. Yamamoto, R. P. Hausinger and J. M. Tiedje (2002). "tfdA-like genes in 2, 4-dichlorophenoxyacetic acid-degrading bacteria belonging to the Bradyrhizobium-Agromonas-Nitrobacter-Afipia cluster in α-Proteobacteria." Applied and environmental microbiology **68**(7): 3449-3454.

Jain, R. K., J. H. Dreisbach and J. C. Spain (1994). "Biodegradation of p-nitrophenol via 1, 2, 4-benzenetriol by an *Arthrobacter* sp." Applied and Environmental Microbiology **60**(8): 3030-3032.

Ka, J., W. E. Holben and J. M. Tiedje (1994). "Genetic and phenotypic diversity of 2, 4-dichlorophenoxyacetic acid (2, 4-D)-degrading bacteria isolated from 2, 4-D-treated field soils." Applied and Environmental Microbiology **60**(4): 1106-1115.

Kitagawa, W., S. Takami, K. Miyauchi, E. Masai, Y. Kamagata, J. M. Tiedje and M. Fukuda (2002). "Novel 2, 4-dichlorophenoxyacetic acid degradation genes from oligotrophic *Bradyrhizobium* sp. strain HW13 isolated from a pristine environment." Journal of bacteriology **184**(2): 509-518.

Konstantinidis, K. T., A. Ramette and J. M. Tiedje (2006). "The bacterial species definition in the genomic era." Philos Trans R Soc Lond B Biol Sci **361**(1475): 1929-1940.

Konstantinidis, K. T. and J. M. Tiedje (2005). "Genomic insights that advance the species definition for prokaryotes." Proceedings of the National Academy of Sciences of the United States of America **102**(7): 2567-2572.

Lanza, V. F., M. de Toro, M. P. Garcillán-Barcia, A. Mora, J. Blanco, T. M. Coque and F. de la Cruz (2014). "Plasmid flux in Escherichia coli ST131 sublineages, analyzed by plasmid constellation network (PLACNET), a new method for plasmid reconstruction from whole genome sequences." PLoS genetics **10**(12): e1004766.

Li, H. and R. Durbin (2009). "Fast and accurate short read alignment with Burrows–Wheeler transform." Bioinformatics **25**(14): 1754-1760.

Liu, H., J.-J. Zhang, S.-J. Wang, X.-E. Zhang and N.-Y. Zhou (2005). "Plasmid-borne catabolism of methyl parathion and p-nitrophenol in *Pseudomonas* sp. strain WBC-3." Biochemical and biophysical research communications **334**(4): 1107-1114.

Lory, S. (2014). The Family Legionellaceae. The Prokaryotes, Springer**:** 387-389.

Love, M. I., W. Huber and S. Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biol **15**(12): 550.

Luo, C., L. M. Rodriguez-R and K. T. Konstantinidis (2014). "MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences." Nucl. Acids. Res.: In press.

Luo, C., D. Tsementzi, N. C. Kyrpides and K. T. Konstantinidis (2012). "Individual genome assembly from complex community short-read metagenomic datasets." The ISME journal **6**(4): 898-901.

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan and Y. Liu (2012). "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler." Gigascience **1**(1): 18.

Madyastha, K. and G. Sridhar (1998). "A novel pathway for the metabolism of caffeine by a mixed culture consortium." Biochemical and biophysical research communications **249**(1): 178-181.

Mazzafera, P., O. Olsson and G. Sandberg (1996). "Degradation of caffeine and related methylxanthines by *Serratia marcescens* isolated from soil under coffee cultivation." Microbial ecology **31**(2): 199-207.

Middelboe, M., N. Jorgensen and N. Kroer (1996). "Effects of viruses on nutrient turnover and growth efficiency of noninfected marine bacterioplankton." Applied and Environmental Microbiology **62**(6): 1991-1997.

Mohanty, S. K., C.-L. Yu, S. Das, T. M. Louie, L. Gakhar and M. Subramanian (2012). "Delineation of the caffeine C-8 oxidation pathway in *Pseudomonas* sp. strain CBB1 via characterization of a new trimethyluric acid monooxygenase and genes involved in trimethyluric acid metabolism." Journal of bacteriology **194**(15): 3872-3882.

Mohapatra, B., N. Harris, R. Nordin and A. Mazumder (2006). "Purification and characterization of a novel caffeine oxidase from *Alcaligenes* species." Journal of biotechnology **125**(3): 319-327.

Musat, N., H. Halm, B. Winterholler, P. Hoppe, S. Peduzzi, F. Hillion, F. Horreard, R. Amann, B. B. Jørgensen and M. M. M. Kuypers (2008). "A single-cell view on the ecophysiology of anaerobic phototrophic bacteria." Proceedings of the National Academy of Sciences **105**(46): 17861-17866.

Nelson, K. E., I. T. Paulsen, J. F. Heidelberg and C. M. Fraser (2000). "Status of genome projects for nonpathogenic bacteria and archaea." Nature biotechnology **18**(10): 1049-1054.

Nielsen, T. K., Z. Xu, E. Gözdereliler, J. Aamand, L. H. Hansen and S. R. Sørensen (2013). "Novel insight into the genetic context of the cadAB genes from a 4-chloro-2-methylphenoxyacetic acid-degrading *Sphingomonas*."

Oh, S., A. Caro-Quintero, D. Tsementzi, N. DeLeon-Rodriguez, C. Luo, R. Poretsky and K. T. Konstantinidis (2011). "Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem." Applied and Environmental Microbiology **77**(17): 6000-6011.

Overbeek, R., T. Begley, R. M. Butler, J. V. Choudhuri, H.-Y. Chuang, M. Cohoon, V. de Crécy-Lagard, N. Diaz, T. Disz and R. Edwards (2005). "The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes." Nucleic acids research **33**(17): 5691-5702.

Payet, J. P. and C. A. Suttle (2013). "To kill or not to kill: The balance between lytic and lysogenic viral infection is driven by trophic status." Limnology and Oceanography **58**(2): 465-474.

Pedrós-Alió, C. (2012). "The rare bacterial biosphere." Annual review of marine science **4**: 449-466.

Prach, K. and L. R. Walker (2011). "Four opportunities for studies of ecological succession." Trends in Ecology & Evolution **26**(3): 119-123.

Quince, C., A. Lanzén, T. P. Curtis, R. J. Davenport, N. Hall, I. M. Head, L. F. Read and W. T. Sloan (2009). "Accurate determination of microbial diversity from 454 pyrosequencing data." Nature methods **6**(9): 639-641.

Ram, A. S. P. and T. Sime-Ngando (2008). "Functional responses of prokaryotes and viruses to grazer effects and nutrient additions in freshwater microcosms." The ISME Journal **2**(5): 498-509.

Ritalahti, K. M., B. K. Amos, Y. Sung, Q. Wu, S. S. Koenigsberg and F. E. Löffler (2006). "Quantitative PCR targeting 16S rRNA and reductive dehalogenase genes simultaneously monitors multiple *Dehalococcoides* strains." Applied and Environmental Microbiology **72**(4): 2765-2774.

Rodriguez-R, L. M. and K. T. Konstantinidis (2014). "Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets." Bioinformatics **30**(5): 629-635.

Rodriguez-R. L-M, Overholt W. A., Hagan C., Huettel M, K. J. E. and K. K. T. (2015). "Microbial community successional patterns in beach sands impacted by the Deepwater Horizon oil spill." The ISME Journal: In press.

Rodriguez-Valera, F., A.-B. Martin-Cuadrado, B. Rodriguez-Brito, L. Pašić, T. F. Thingstad, F. Rohwer and A. Mira (2009). "Explaining microbial population genomics through phage predation." Nature Reviews Microbiology **7**(11): 828-836.

Sakai, Y., N. Ogawa, Y. Shimomura and T. Fujii (2014). "A 2, 4-dichlorophenoxyacetic acid degradation plasmid pM7012 discloses distribution of an unclassified megaplasmid group across bacterial species." Microbiology **160**(Pt 3): 525-536.

Schäfer, A., H. Harms and A. J. Zehnder (1996). "Biodegradation of 4-nitroanisole by two *Rhodococcus* spp." Biodegradation **7**(3): 249-255.

Shade, A., S. E. Jones, J. G. Caporaso, J. Handelsman, R. Knight, N. Fierer and J. A. Gilbert (2014). "Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity." MBio **5**(4): e01371-01314.

Sogin, M. L., H. G. Morrison, J. A. Huber, D. Mark Welch, S. M. Huse, P. R. Neal, J. M. Arrieta and G. J. Herndl (2006). "Microbial diversity in the deep sea and the underexplored "rare biosphere"." Proc Natl Acad Sci U S A **103**(32): 12115-12120.

Spain, J. C. and D. T. Gibson (1991). "Pathway for biodegradation of p-nitrophenol in a *Moraxella* sp." Applied and Environmental Microbiology **57**(3): 812-819.

Stanier, R. (1942). "The Cytophaga group: a contribution to the biology of myxobacteria." Bacteriological reviews **6**(3): 143.

Stanier, R. Y., N. J. Palleroni and M. Doudoroff (1966). "The aerobic pseudomonads a taxonomic study." Microbiology **43**(2): 159-271.

Stibal, M., J. Bælum, W. E. Holben, S. R. Sørensen, A. Jensen and C. S. Jacobsen (2012). "Microbial degradation of 2, 4-dichlorophenoxyacetic acid on the Greenland ice sheet." Applied and environmental microbiology **78**(15): 5070-5076.

Stoeck, T. and S. Epstein (2009). "Protists and the rare biosphere. Crystal Ball." Environ Microbiol Reports **1**: 3-26.

Su, X., W. Pan, B. Song, J. Xu and K. Ning (2014). "Parallel-META 2.0: enhanced metagenomic data analysis with functional annotation, high performance computing and advanced visualization." PloS one **9**(3): e89323.

Sullivan, M. J., N. K. Petty and S. A. Beatson (2011). "Easyfig: a genome comparison visualizer." Bioinformatics **27**(7): 1009-1010.

Summers, R. M., T. M. Louie, C.-L. Yu, L. Gakhar, K. C. Louie and M. Subramanian (2012). "Novel, highly specific N-demethylases enable bacteria to live on caffeine and related purine alkaloids." Journal of bacteriology **194**(8): 2041-2049.

Summers, R. M., S. K. Mohanty, S. Gopishetty and M. Subramanian (2015). "Genetic characterization of caffeine degradation by bacteria and its potential applications." Microbial biotechnology **8**(3): 369-378.

Summers, R. M., J. L. Seffernick, E. M. Quandt, C. L. Yu, J. E. Barrick and M. V. Subramanian (2013). "Caffeine junkie: an unprecedented glutathione S-transferase-dependent oxygenase required for caffeine degradation by *Pseudomonas putida* CBB5." Journal of bacteriology **195**(17): 3933-3939.

Thingstad, T. and R. Lignell (1997). "Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand." Aquatic Microbial Ecology **13**(1): 19-27.

Tonso, N., V. Matheson and W. Holben (1995). "Polyphasic characterization of a suite of bacterial isolates capable of degrading 2, 4-D." Microbial ecology **30**(1): 3-24.

Top, E. M., W. E. Holben and L. J. Forney (1995). "Characterization of diverse 2, 4-dichlorophenoxyacetic acid-degradative plasmids isolated from soil by complementation." Applied and Environmental Microbiology **61**(5): 1691-1698.

Torsvik, V., J. Goksoyr and F. L. Daae (1990). "High diversity in DNA of soil bacteria." Appl Environ Microbiol **56**(3): 782-787.

Trefault, N., R. De la Iglesia, A. M. Molina, M. Manzano, T. Ledger, D. Pérez‐Pantoja, M. A. Sánchez, M. Stuardo and B. González (2004). "Genetic organization of the catabolic plasmid pJP4 from *Ralstonia eutropha* JMP134 (pJP4) reveals mechanisms of adaptation to chloroaromatic pollutants and evolution of specialized chloroaromatic degradation pathways." Environmental microbiology **6**(7): 655-668.

Turner, S., K. M. Pryer, V. P. Miao and J. D. Palmer (1999). "Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis1." Journal of Eukaryotic Microbiology **46**(4): 327-338.

Wang, Q., G. M. Garrity, J. M. Tiedje and J. R. Cole (2007). "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy." Applied and environmental microbiology **73**(16): 5261-5267.

Weitz, J. S. and S. W. Wilhelm (2012). "Ocean viruses and their effects on microbial communities and biogeochemical cycles." F1000 biology reports **4**.

Wells, L. E. and J. W. Deming (2006). "Significance of bacterivory and viral lysis in bottom waters of Franklin Bay, Canadian Arctic, during winter." Aquatic microbial ecology **43**(3): 209-221.

Whitman, W. B., D. C. Coleman and W. J. Wiebe (1998). "Prokaryotes: the unseen majority." Proceedings of the National Academy of Sciences **95**(12): 6578-6583.

Wu, C. H., R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang and R. Lopez (2006). "The Universal Protein Resource (UniProt): an expanding universe of protein information." Nucleic acids research **34**(suppl 1): D187-D191.

Yamaoka-Yano, D. and P. Mazzafera (1998). "Degradation of caffeine by *Pseudomonas putida* isolated from soil." Allelopathy J **5**(1): 23-34.

Yu, C. L., Y. Kale, S. Gopishetty, T. M. Louie and M. Subramanian (2008). "A novel caffeine dehydrogenase in *Pseudomonas* sp. strain CBB1 oxidizes caffeine to trimethyluric acid." Journal of bacteriology **190**(2): 772-776.

Yu, C. L., T. M. Louie, R. Summers, Y. Kale, S. Gopishetty and M. Subramanian (2009). "Two distinct pathways for metabolism of theophylline and caffeine are coexpressed in *Pseudomonas putida* CBB5." Journal of bacteriology **191**(14): 4624-4632.

Yu, C. L., R. M. Summers, Y. Li, S. K. Mohanty, M. Subramanian and R. M. Pope (2014). "Rapid identification and quantitative validation of a caffeine-degrading pathway in *Pseudomonas* sp. CES." Journal of proteome research **14**(1): 95-106.

Zerbino, D. R. and E. Birney (2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." Genome research **18**(5): 821-829.

Zhang, J., K. Kobert, T. Flouri and A. Stamatakis (2014). "PEAR: a fast and accurate Illumina Paired-End reAd mergeR." Bioinformatics **30**(5): 614-620.

Zhang, J.-J., H. Liu, Y. Xiao, X.-E. Zhang and N.-Y. Zhou (2009). "Identification and characterization of catabolic para-nitrophenol 4-monooxygenase and para-benzoquinone reductase from *Pseudomonas* sp. strain WBC-3." Journal of bacteriology **191**(8): 2703-2710.

Zhang, S., W. Sun, L. Xu, X. Zheng, X. Chu, J. Tian, N. Wu and Y. Fan (2012). "Identification of the para-nitrophenol catabolic pathway, and characterization of three enzymes involved in the hydroquinone pathway, in *Pseudomonas* sp. 1-7." BMC microbiology **12**(1): 27.

Zhou, J., Y. Deng, P. Zhang, K. Xue, Y. Liang, J. D. Van Nostrand, Y. Yang, Z. He, L. Wu and D. A. Stahl (2014). "Stochasticity, succession, and environmental perturbations in a fluidic ecosystem." Proceedings of the National Academy of Sciences **111**(9): E836-E845.